FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Conditional Diffusion Models for Visual Anonymization of Medical Case-based Explanations

Filipe Pinto Campos

Mestrado em Engenharia Informática e Computação

Supervisor: Luís Filipe Teixeira

Co-Supervisor: Wilson Silva

July 23, 2024

Conditional Diffusion Models for Visual Anonymization of Medical Case-based Explanations

Filipe Pinto Campos

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Augusto de Sousa
Referee: Prof. João Neves
Referee: Prof. Luís Filipe Teixeira

July 23, 2024

Resumo

Técnicas de *Deep Learning* são capazes de melhorar a eficiência do diagnóstico médico, desafiando a precisão de peritos humanos. No entanto, existe uma falta de confiança nos métodos de tomada de decisão baseados em DL devido à sua falta de interpretabilidade. *Case-based explanations*, ou explicações baseadas em casos auxiliam na compreensão desses métodos, apresentando, para uma decisão específica, imagens semelhantes ou prototípicas que partilham o mesmo diagnóstico ou destacam diferenças visuais entre vários diagnósticos. No entanto, estas imagens podem conter dados privados, sujeitos a regulamentações rigorosas, como o Regulamento Geral de Proteção de Dados (RGPD) e a Lei de Portabilidade e Responsabilidade do Seguro de Saúde (HIPAA), limitando significativamente a sua utilidade. Esta questão é particularmente importante em *Federated Learning*, onde algoritmos são treinados colaborativamente entre várias instituições, e onde evitar a partilha de dados identificáveis é crucial. A anonimização surge como uma solução para este problema, removendo características pessoais dos dados e permitindo a sua partilha sem tais restrições.

Os métodos estado-da-arte, baseados em GANs, geram explicações com baixa qualidade visual. Nesta dissertação exploramos novas abordagens baseadas em *conditional diffusion models*, usando modelos inovadores ou combinando um sistema de anonimização e modelos generativos existentes. Também exploramos métodos centralizados e baseados em Federated Learning para obter as imágens sintéticas e anónimas geradas por estes métodos.

Validamos as nossas propostas em tarefas de radiografia torácica pois são uma das modalidades de imagem médica mais utilizadas. Especificamente, realizamos experiências nos *datasets* MIMIC-CXR-JPG, BRAX e CheXpert. Avaliamos os métodos de geração de explicações relativamente a vários critérios, como o realismo das imagens, utilidade e privacidade das explicações geradas. Além disso, avaliamos os nossos métodos de recuperação com a ajuda de uma radiologista experiente, avaliando tanto a relevância das explicações recuperadas em comparação com as recuperadas por um especialista como a concordância entre as *labels* das imagens geradas e os seus diagnósticos. Neste trabalho conseguimos um método capaz de criar e recuperar explicações sintéticas baseadas em casos que são realistas e anónimas, e tão relevantes na explicação de decisões médicas quanto imagens reais, com base na avaliação mencionada do especialista.

Esta dissertação esforça-se por garantir que os profissionais de saúde possam partilhar explicações baseadas em casos de maneira segura e livre, sem violar o direito essencial à privacidade do paciente. Por sua vez, contribui para a adopção mais ampla de métodos de aprendizagem profunda (DL) no campo médico e aborda preocupações críticas relacionadas com a privacidade de dados e conformidade com os quadros regulatórios.

Palavras-chave: Deep Learning, Interpretabilidade, Privacidade, Modelos de Difusão, Aprendizagem Federada

Abstract

Deep-learning techniques can improve the efficiency of medical diagnosis while challenging human experts' accuracy. However, there is a lack of trust in Deep Learning based decision-making methods due to their lack of interpretability. Case-based explanations help comprehend these methods by presenting, for a specific classification, similar or prototypical images sharing the same diagnosis or by highlighting visual differences between various diagnoses. Yet, these may contain private data, subject to strict regulations, such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA), heavily limiting their utility. This issue is particularly important in Federated Learning, where algorithms are trained collaboratively between multiple institutions, but avoiding sharing identifiable data is paramount. Anonymization serves as the solution to this problem, removing personal characteristics from data and thus enabling its shareability without such constraints.

The current state-of-the-art methods, based on GANs, generate explanations with low-quality images. In this dissertation, we explore new approaches based on conditional diffusion models, using either novel models or by coupling an anonymization system to existing generative models. We also explore centralized and Federated-Learning approaches to retrieve the anonymous synthetic images generated using these methods.

We validate our proposals on chest X-ray tasks because they are one of the most medical imaging modalities. Specifically, we perform experiments on the MIMIC-CXR-JPG, BRAX and CheXpert datasets. We evaluate the explanation generation methods regarding multiple criteria such as image realism, utility and privacy of the generated explanations. Additionally, we evaluate our retrieval methods with the aid of an experienced radiologist, evaluating both the relevance of the retrieved explanations compared to the ones retrieved by an expert and the label agreement between generated images and their diagnosis. We achieve a method capable of creating and retrieving synthetic case-based explanations that are realistic and anonymous and as relevant in explaining medical decisions as real images based on the aforementioned expert evaluation.

This work strives to ensure that healthcare professionals can securely and freely share case-based explanations without violating the patient's essential right to privacy. In turn, it contributes to the broader adoption of DL methods in the medical field and addresses critical concerns related to data privacy and compliance with regulatory frameworks.

Keywords: Deep Learning, Interpretability, Privacy-preserving Machine Learning, Diffusion Models, Federated-Learning

Acknowledgments

First and foremost, I extend my gratitude to my supervisors, Professors Luís Teixeira and Wilson Silva. Thank you for all the support and guidance, for all the discussions and encouragement, which forced me to strive for better. Most importantly, thank you for putting up with me through it all; it meant a lot.

I am grateful to INESC-TEC, the VCMI group and the CAGING project under which I received a research grant. This has both financially supported and upped the ante of this work, encouraging me to strive towards higher ambitions.

To my colleagues in the AI Technology for Life group, thank you for the stimulating discussions and support. Being surrounded by you all has made me learn a lot and realize how much I still have to learn; I loved it. I will never forget our coffee breaks/expeditions, the group lunches and how you all made me feel welcome; this experience was truly memorable and enjoyable.

To everyone I have met at the NKI, thank you for the delightful Friday cake meetings and *borrels*, which, though infrequent, were always amazing. Being present gave me a lot of insight into the medical world and its apparently endless drama.

To the professors at FEUP who, throughout these years, have helped me grow as a student, I would not be where I am today without you, and for that, I am thankful. I hope you continue to guide and help even more students in the times to come; your impact is immeasurable.

To the friends I've had the pleasure of meeting this semester, thank you for all the incredible moments and sleepless nights and for encouraging me to step out of my comfort zone (often whilst on a bike). Who knew it was possible to compress so many experiences in a five-month period while simultaneously feeling the time rushing by with incredible haste? For that, I am grateful to you all.

To my long-time friends, thank you for your unwavering support over the past five years. I know it has been a long journey, but you always helped me push through, even through the toughest times. Thank you also for being there for the good times as well, which were equally important. And during these recent times, despite the 1600km distance, you have made me feel like I was right at home as usual, like I can just step back into your company and carry on as if nothing has changed. I couldn't wish for anything better.

To my parents, I am grateful for the support you have given me throughout these years; without you, this milestone would have been unattainable. To my sister, thank you for always being annoying and able to be annoyed; somehow, you always manage to make the dull days more remarkable. Lastly, to my dogs, Kikas and Holly, for always being there; your company is a constant delight.

This work is financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project CAGING, with reference 2022.10486.PTDC (DOI 10.54499/2022.10486.PTDC).

Filipe Pinto Campos

"And so, does the destination matter? Or is it the path we take? I declare that no accomplishment has substance nearly as great as the road used to achieve it. We are not creatures of destinations. It is the journey that shapes us. Our callused feet, our backs strong from carrying the weight of our travels, our eyes open with the fresh delight of experiences lived."

Brandon Sanderson

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Motivation . . . 1
- 1.3 Objectives . . . 2
- 1.4 Main Contributions . . . 2
- 1.5 Document Structure . . . 3

2 Background **4**
- 2.1 Overview . . . 4
- 2.2 Anonymization . . . 4
- 2.3 Generative Models . . . 6
  - 2.3.1 Variational Autoencoders . . . 6
  - 2.3.2 Generative Adversarial Networks . . . 7
  - 2.3.3 Diffusion Probabilistic Models . . . 7
  - 2.3.4 Denoising Diffusion Implicit Models . . . 10
- 2.4 Federated Learning . . . 11
- 2.5 Summary . . . 12

3 Literature Review **14**
- 3.1 Overview . . . 14
- 3.2 Diffusion Models . . . 15
  - 3.2.1 Classifier-Guidance in Diffusion Models . . . 15
  - 3.2.2 Latent Diffusion Models . . . 16
  - 3.2.3 Medical Image Generation with Diffusion Models . . . 17
  - 3.2.4 Semantic Latent Manipulation in Diffusion Models . . . 18
- 3.3 Visual Anonymization . . . 20
  - 3.3.1 Classic Methods . . . 21
  - 3.3.2 Deep Learning Based Methods . . . 21
- 3.4 Interpretability . . . 23
  - 3.4.1 Case-Based Interpretability . . . 24
- 3.5 De-Anonymization Methods . . . 27
- 3.6 Summary . . . 30

4 Preliminary Data Choices **32**
- 4.1 Overview . . . 32
- 4.2 Datasets . . . 32
- 4.3 Task choice . . . 33
- 4.4 Summary . . . 34

5 In-model Anonymization **35**
5.1 Overview . . . 35
5.2 Differentially Private Diffusion Model . . . 35
5.2.1 Experiments . . . 36
5.3 Decoder-level Anonymization . . . 36
5.3.1 Method . . . 38
5.3.2 Experiments . . . 39
5.4 Classifier-free Identity Guidance for Privacy-Preserving Image Generation . . . 40
5.4.1 Method . . . 40
5.4.2 Experiments . . . 42
5.5 Summary . . . 44

6 Post-model Anonymization **45**
6.1 Overview . . . 45
6.2 Method . . . 45
6.2.1 Image Generation using Latent Diffusion Models . . . 46
6.2.2 Retrieval and Verification Network . . . 46
6.2.3 Anonymization Pipeline . . . 47
6.3 Experiments . . . 48
6.4 Results and discussion . . . 48
6.5 Summary . . . 50

7 Case-based Explanation Retrieval **51**
7.1 Overview . . . 51
7.2 Retrieval in a Centralized setting . . . 51
7.2.1 Method . . . 51
7.2.2 Experiments . . . 51
7.2.3 Results and discussion . . . 52
7.3 Retrieval in a Federated-Learning setting . . . 54
7.3.1 Method . . . 55
7.3.2 Experiments . . . 55
7.3.3 Results and discussion . . . 56
7.4 Practical Considerations . . . 56
7.4.1 Model Usage . . . 57
7.4.2 Safety . . . 60
7.5 Summary . . . 61

8 Conclusion **62**

References **64**

List of Figures

2.1 High-level overview of architectures of commonly used generative models 6
2.2 Illustration of the reverse diffusion model process for generating images from Gaussian noise 8
2.3 Federated Learning Framework. 12

3.1 Latent Diffusion Model architecture. 16
3.2 Overview of manipulation approaches for diffusion models 18
3.3 Architecture of a Diffusion Autoencoder 19
3.4 Examples of applying K-Same and K-Same-Select 21
3.5 Architecture proposed by Packhäuser for the generation of anonymous chest radiographs 23
3.6 Pipeline for the DP-FL model 25
3.7 Overview of the privacy-preserving model proposed by Montenegro *et al.* 26
3.8 BPMN diagram of the proposed workflow for a practical application of the method under study 28
3.9 Diagram of membership inference attack in a black-box setting 29

4.1 Comparison between a healthy X-ray and one that presents Cardiomegaly. 34

5.1 In-model approach: generate anonymous synthetic images and retrieve them. . . 35
5.2 Samples generated using the DPDM model. 37
5.3 Overview of the training process for the decoder-level anonymization method. . . 37
5.4 Comparison between input and output images using VAE trained using a privacy loss function 40
5.5 Overview of the sampling process for the classifier-free identity guidance mechanism. 42
5.6 Generation of images using the classifier-free guidance mechanism for two conditions, Cardiomegaly and Pleural Effusion. 43
5.7 Generation of images using the classifier-free guidance mechanism for two conditions, Cardiomegaly and an identity embedding. 43

6.1 Post-model approach: first, generate synthetic images, then, anonymize them and retrieve them. 45
6.2 Anonymization Pipeline. 46
6.3 Different image pairs of anonymous images and their closest real counterpart in the training set, sorted by the predicted likelihood of belonging to the same patient. 50

7.1 Boxplots regarding nDCG for two retrieval approaches, CNN and SSIM for the Top-4, Top-7 and Top-10 retrieved images. 53

7.2 Example test image and corresponding Top-4 catalogue images retrieved by each corresponding method. 54
7.3 Overview of the Federated-learning architecture used. 55
7.4 Boxplots regarding nDCG for three retrieval approaches, CNN, CNN trained using FL and SSIM for the Top-4, Top-7 and Top-10 retrieved images. 57
7.5 Example of images retrieved using our FL retrieval model as explanations for multiple test images. 58
7.6 Aggregation mechanism to build a centralized embedding database. 59
7.7 Example of a retrieval mechanism under a federated learning setting. 59

List of Tables

3.1 Overview of attack methods proposed in Literature. 30

4.1 Chest X-ray dataset comparison. 32

5.1 Results of training VAE with privacy loss function. 39

6.1 Quantitative evaluation of the images generated using the Medfusion model. . . . 48
6.2 Classification performance reported for both the MIMIX-CXR-JPG and a synthetic dataset. 49
6.3 Quantitative evaluation of patient verification model. 49
6.4 Quantitative evaluation of patient retrieval model. 49

7.1 Architecture of the DenseNet-121 model. 52
7.2 Label agreement between expert evaluation and the ground-truth label. 54
7.3 Performance of the DenseNet121 model on a centralized setting. 56
7.4 Comparison in performance, measured using AUC, of models trained on different datasets and a Federated-Learning approach. 56
7.5 Overview of possible attack methods for different usage settings for the Case-based explanation retrieval system. 60

Abbreviations

AUC	Area under the ROC Curve.
CNN	Convolutional Neural Network.
CT	Computarized Tomography.
DDIM	Denoising Diffusion Implicit Models.
DDPM	Denoising Probabilistic Diffusion Models.
DL	Deep Learning.
DNN	Deep Neural Network.
DPDM	Differentially Private Diffusion Models.
FID	Fréchet Inception Distance.
FL	Federated Learning.
GAN	Generative Adversarial Network.
GDPR	General Data Protection Regulation.
HIPAA	Health Insurance Portability and Accountability Act.
KL	Kullback-Leibler divergence.
LDM	Latent Diffusion Model.
LPIPS	Learned Perceptual Image Patch Similarity.
MIA	Membership Inference Attack.
ML	Machine Learning.
MRI	Magnetic Resonance Imaging.
MSE	Mean Squared Error.
NN	Neural Network.
PA	Posteroanterior.
PET	Positron Emission Tomography.
SSIM	Structural Similarity Index.
VAE	Variational Autoencoder.
XAI	Explainable Artificial Intelligence.

Chapter 1

Introduction

1.1 Context

Medical imaging stands as a cornerstone in diagnosing various diseases, employing techniques such as X-ray or MRI to detect multiple conditions. These imaging modalities play an important role in providing clinicians with insights into a patient's health, facilitating accurate and timely diagnoses.

In recent years, the integration of Deep Learning (DL) has marked a shift in medical diagnostics, showcasing capabilities that can often rival or even surpass human performance. The advent of DL algorithms in analyzing medical images has significantly enhanced the accuracy and efficiency of diagnostic processes. However, this leap in computational prowess brings with it a profound challenge – the inherent lack of interpretability in Deep Learning models.

While these models exhibit remarkable diagnostic capabilities, their decision-making processes remain largely opaque, posing a barrier to understanding the rationale behind their predictions. This lack of interpretability raises legitimate concerns, especially in the context of critical medical decisions where transparency is necessary for establishing trust and facilitating effective collaboration between machine intelligence and healthcare professionals.

1.2 Motivation

To permit the integration of DL in the workflow of hospitals and respective doctors, we must first provide methods to glean insights about the decision-making mechanisms of models. One such way is through the use of case-based explanations, which explain by example, closely mimicking the rationale of human professionals. Using these mechanisms, a doctor would have access to a prediction and a set of image explanations for each machine-made diagnosis. These can be similar cases with the same diagnosis, prototypical images, or counterfactual explanations highlighting the differences between images that contain and do not contain the diagnosis.

These explanations ideally should be shareable between medical professionals, between doctor and patient or even between hospitals, which is particularly relevant when exemplifying rare

conditions. Yet, this is not possible since medical images contain personal identification, which may allow for the re-identification of patients; therefore, they are protected under strict regulations. These issues can be overcome by anonymizing said explanations, therefore protecting the privacy of patients.

1.3 Objectives

This work aims to improve the explainability of Deep-learning-based medical decision systems and, consequently, improve the efficiency and accuracy of diagnosis in medical institutions. To do so, we propose to develop a system that generates anonymous case-based explanations using diffusion models and retrieves them, either in a centralized or decentralized manner, as a mechanism to explain a classifier's decisions. Additionally, due to the sensitive nature of medical data, this work endeavours to evaluate and quantify the risk of exposing the identity of patients when sharing anonymized images.

Formally, these objectives can be condensed into the following set of research questions (RQ):

RQ1. Can anonymous case-based explanations be generated using diffusion models?

RQ2. Is it possible to retrieve the aforementioned explanations in a federated learning setting?

RQ3. Can the de-anonymization risk associated with the aforementioned anonymous explanations be evaluated and quantified?

After answering these questions, we expect to have a system that generates, anonymizes, and retrieves case-based explanations either in centralized or Federated-Learning settings. Additionally, we should be able to measure the risk of de-anonymization so responsible entities such as hospitals can be aware of the risks associated with the approach and make better-informed decisions.

1.4 Main Contributions

The main contributions of this dissertation are the following:

1. We investigate three distinct approaches for in-model image anonymization using diffusion models and how their utility as case-based explanations.
2. We propose an image anonymization method, post-model image anonymization, which employs a latent diffusion model in a three-step procedure: generating a catalogue of synthetic images, removing the images that closely resemble existing patients, and using this anonymous catalogue during an explanation retrieval process.
3. We explore how to retrieve case-based explanations in a centralized setting and in a Federated-Learning setting.

4. We evaluate the proposed methods on the MIMIC-CXR-JPG [41] dataset for the centralized experiments and additionally on BRAX [83] and CheXpert [35] datasets for the Federated-Learning experiments. Additionally, the quality of the retrieved case-based explanations was evaluated by an experienced radiologist.

Additionally, we published a research paper [6] for the EXPLIMED workshop within ECAI 2024 which covers the proposed post-model image anonymization approach, the centralized retrieval of case-based explanations and its evaluation.

1.5 Document Structure

The remainder of this document is structured as follows. In Chapter 2, we present the background concepts which serve as the foundation for the rest of the dissertation. In Chapter 3, we analyze studies which tackle problems adjacent to our work and help define a proposed solution for our problem. Chapter 4 presents the datasets and tasks chosen to evaluate the methods presented in the following chapters. In Chapter 5, we explore methods to create a model which generates anonymized images while Chapter 6 presents a two-stage approach where a model first generates synthetic images and; afterwards, a secondary system is used to identify images which closely resemble existing patients and removes them. Chapter 7 follows up on the previous two chapters by presenting methods to retrieve anonymous images and use them as case-based explanations in a centralized setting or a federated learning setting where multiple hospitals collaboratively train a retrieval model and share image catalogues. Finally, in Chapter 8, we draw concluding remarks and outline future work.

Chapter 2

Background

2.1 Overview

This chapter endeavours to present the background knowledge which serves as the foundation of this work. In Section 2.2, we introduce the concept of anonymization, which is of utmost importance in this work. In Section 2.3, we present an overview of the existing landscape of generative models and delve into the details of diffusion models, which will be the main focus of this work. Section 2.4 will present the foundations of Federated Learning, which is important to permit the sharing of case-based explanations across hospitals or institutions without leaking sensitive data. Finally, in Section 2.5, we present a short summary and closing remarks about the contents of this chapter.

2.2 Anonymization

Many countries are establishing and developing ever stricter data protection laws to preserve the citizen's fundamental right to privacy. Although important, these regulations, such as the European General Data Protection Regulation (GDPR) [12] or the Health Insurance Portability and Accountability Act (HIPAA) [100] in the United States of America slow down advances in medical machine learning applications by imposing restrictions on data sharing, making data a scarce resource. Anonymization stands as a prominent way of enabling data sharing without any of the aforementioned restrictions because it allows for the removal of any personally identifiable information, making de-anonymization near-impossible.

The GDPR provides definitions of two important concepts, **pseudonymisation** and **anonymization**. The former ensures that the personal data is unidentifiable without the use of some external data. This second clause implies it is susceptible to cross-referencing attacks, which combine information from multiple data sources. In opposition, anonymization must guarantee that the data is irreversibly non-identifiable regardless of the circumstances. That is, a datum is only anonymized if it is impossible to de-anonymize, considering all the reasonable methods that an attacker may use to attempt such a feat.

Generally, a way of analyzing the degree of anonymization of a dataset is by identifying whether or not a data point can be narrowed down to a single individual; this is also known as *singling out*. This is taken under consideration in Recital 26, Article 4 of the GDPR where The European Parliament and Council of the European Union specify:

> "The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly."

Therefore, any proposed method of anonymization must undergo a careful study to identify all the reasonably likely methods to de-anonymize the data it anonymizes. After this careful process, the data is no longer susceptible to the same data protection principles identified under the GDPR, as specified in the same Recital:

> "The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

Analyzing these concepts in the domain of this work, we must consider that seemingly innocuous data, such as medical exams, can be used to single out individuals, spurring the need to anonymize every image we intend to share. In fact, different types of scans have been shown to be capable of identifying the individuals they belong to [87, 99], including Magnetic resonance imaging (MRI) [38], computerized tomography (CT) scans [81] or positron emission tomography (PET) scans [88].

Guaranteeing full anonymity of data is a difficult process since it involves protecting against multiple attack vectors which are ever evolving. Given this, Seastedt *et al.* [89] argues that there is a clear trade-off between minimising the risk of identity exposure and the usefulness of the data. Evidently, ceasing all data-sharing activities would be the most effective means to prevent privacy issues. However, such an approach could significantly impede progress in a field where advancements could directly translate to benefits to healthcare systems. Consequently, new anonymisation methods must make compromises and require careful analysis of the risks they carry before being deployed in sensitive scenarios.

2.3 Generative Models

Generative models pertain to a family of ML models whose goal, as the name implies, is to generate new data points based on existing training data. This generation typically takes into consideration key factors such as realism and diversity.

For this work, generative models play a key role, since the most prominent avenue to perform anonymization by either visually modifying already existing pictures or by generating completely new images whose identity cannot be traced back to a specific patient.

This section will introduce a set of architectures prominent in the image generation field. In Fig. 2.1, we present an overview of the architectures we will summarise.

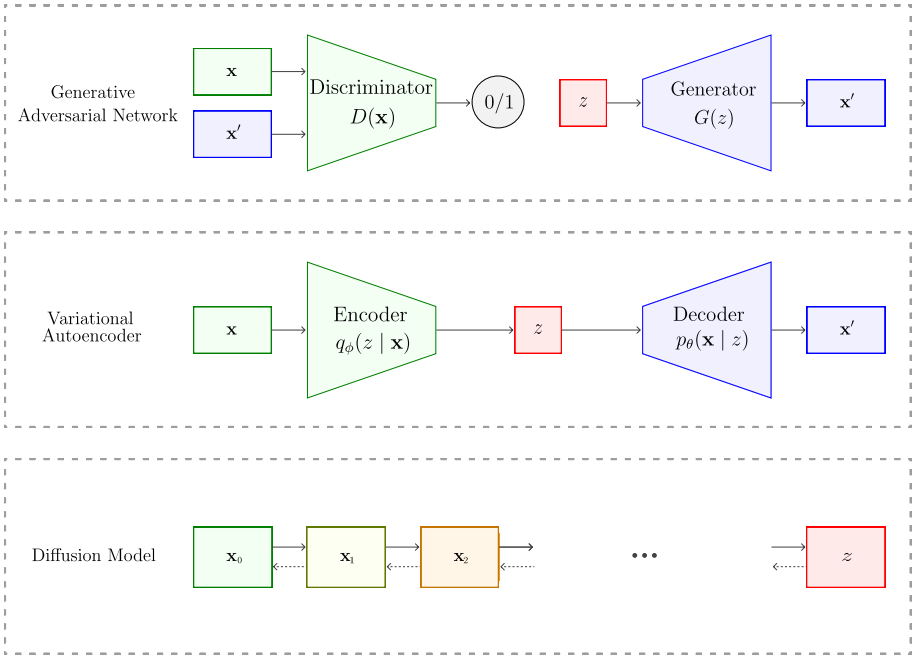


Figure 2.1: High-level overview of architectures of commonly used generative models [105]. Generative Adversarial Networks are composed of two adversarial networks, a generator G, which creates synthetic samples and a discriminator D, which predicts whether a sample x is real or synthetic. Variational autoencoders capture the original data distribution and represent it using $q_\theta(z \mid x)$. Afterwards, a latent sampled from this distribution is decoded back into the original image space. Diffusion models follow an iterative process where stochastic noise is progressively added to an original image x_0 until the image becomes pure Gaussian noise z. Then the network reverts this by iteratively predicting and removing the noise which was added at each timestep, therefore generating a new image.

2.3.1 Variational Autoencoders

Autoencoders are networks comprised of two neural networks: an encoder and a decoder. The encoder performs dimensionality reduction, compressing an input image into a much denser latent space. The decoder performs the reverse process, reconstructing this latent space into the original image.

A Variational Autoencoder (VAE) [51, 50, 104] maps an input x into a distribution p_θ, parameterized by θ. Afterwards, a latent encoding vector z is decoded back into the original data space $\hat{x}$.

The probability $p_\theta(x \mid z)$ defines a *probabilistic decoder* while the approximation function $q_\theta(z \mid x)$ is the *probabilistic encoder*.

To optimize these models, the Kullback-Leibler divergence (KL) is used. The KL divergence $D_{\mathrm{KL}}(X \,||\, Y)$ measures how much information is lost if we represent distribution X using distribution Y. In an VAE the posterior $q_\theta(z \mid x)$ should be close to the real one $p_\theta(z \mid x)$, therefore we pretend to minimize $D_{\mathrm{KL}}(q_\theta(z \mid x) \mid p_\theta(z \mid x))$. The loss function used to train these models is illustrated in Equation 2.1.

$$\mathcal{L}_{\mathrm{VAE}} = -\mathbb{E}_{z\sim q_\theta(z|x)}\left[\log p_\theta(x \mid z) + D_{\mathrm{KL}}\left(q_\theta(z \mid x) \,||\, p_\theta(z)\right)\right] \tag{2.1}$$

2.3.2 Generative Adversarial Networks

A Generative Adversarial Network (GAN) [24] consists of two models with distinct objectives which, when combined, collaborate towards generating realistic images. A generator, G, creates synthetic images based on the data distribution z. The discriminator, D, is presented with either real ($D(x)$) or generated ($D(G(z))$ images and is tasked with distinguishing whether or not the presented image is real. In this "competition" the generator must create ever-more-realistic images to keep fooling the discriminator. This process can be formalized as a min-max problem, as shown in equation 2.2.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\mathrm{data}(x)}}\left[\log D(x)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1 - D(G(z))\right)\right] \tag{2.2}$$

This adversarial formulation makes the training process unstable when compared to other methods, which leads to issues such as vanishing and exploding gradient [16] or mode collapse, severely limiting the capability of the model generating samples across the entire underlying data distribution. Regardless of these issues, this family of networks have been extensively used for a plethora of tasks [36], proving versatile and capable of achieving remarkable results.

2.3.3 Diffusion Probabilistic Models

Diffusion models, originally inspired by non-equilibrium thermodynamics [94], are a category of generative models which, in recent years, have been brought to the forefront of Computer Vision research for tasks such as image generation or super-resolution [13]. This large investment in research of diffusion models was mainly stirred by Denoising Diffusion Probabilistic Models (DDPM) [32], which showcased the capability of generating high-quality image samples. Recently these models have been applied to the medical domain for multiple tasks such as image generation, segmentation or reconstruction and for multiple modalities such as MRI, CT, PET or X-ray [46, 47, 58].

Diffusion models have a more stable training process and empirically have shown results more impressive than previous state-of-the-art GAN alternatives. The major downside which encompasses most diffusion models is the slow image generation time because they rely on an iterative process which is described by a Markov Chain [72] (Figure 2.2), although changes have been proposed to dampen this issue [95, 84, 71].

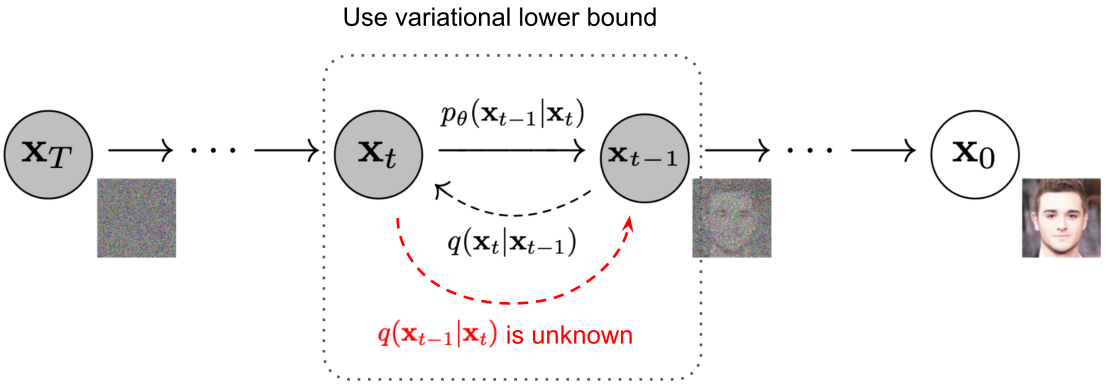


Figure 2.2: Illustration of the reverse diffusion process for generating images from Gaussian noise. Source: [105, 32]

The diffusion process is split into two sub-tasks, the *forward process* and the *reverse process*. During the forward process, Gaussian noise is iteratively added to an image until it becomes pure noise. The reverse process, as the name implies, endeavours to undo these changes by iteratively removing noise until the original image is obtained. These processes are detailed in Subsections 2.3.3.1 and 2.3.3.2, respectively.

2.3.3.1 Forward diffusion process

Consider a data point sampled from the real data distribution $x_0 \sim q(x)$, corresponding to an image sampled from the original training data. During the *forward process*, Gaussian Noise is sequentially added to x_t, where t is the current timestep, at a pre-defined rate set by the *variance schedule* $\{\beta_t \in (0,1)\}_{t=1}^{T}$. A single forward step is defined according to expression 2.3,

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t I), \forall t \in \{1, ..., T\} \tag{2.3}$$

where $\mathcal{N}(x; \mu, \sigma)$ represents a normal distribution of mean μ and covariance σ and I is the identity matrix.

As t increases, the original data point x_0 will become increasingly indistinguishable. As $T \to \infty$, x_T will represent an isotropic Gaussian distribution. A data point composed of pure noise, x_T can be obtained according to equation 2.4 which describes a Markov Chain,

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x \mid x_{t-1}) \tag{2.4}$$

With this definition, we face an immediate problem because to sample an arbitrary x_t, we are required to compute the entire sequence $x_{t-1}, x_{t-2}, ..., x_1$. This issue can be solved by using a re-parameterization trick. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, by modifying equation 2.3 we obtain equation 2.5,

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} \cdot x_0, (1 - \bar{\alpha}_t)I) \tag{2.5}$$

this new formulation allows us to sample an arbitrary x_t based on the original input x_0. In practice, this characteristic is crucial to enable faster training times compared to the initial formulation since it allows us to train models at different timesteps without having to re-calculate the entire noising chain.

2.3.3.2 Reverse diffusion process

During the reverse diffusion process, we aim to calculate $q(x_{t-1} \mid x_t)$ enabling us to recreate a true data sample from pure Gaussian Noise. Yet, this goal is intractable. Therefore we need a model p_θ to approximate the conditional probability distribution, which implies estimating both the mean μ_θ and the variance Σ_θ, where θ are the model parameters. The formalization of a single step is defined in Equation 2.6, while Equation 2.7 summarizes the entire reverse process.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \forall t \in \{1...T\} \tag{2.6}$$

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t) \tag{2.7}$$

In the original implementation [32], the authors decided to fix the variance at a value of $\Sigma_\theta(x_t, t) = \sigma_t^2 I$ and $\sigma_t^2 = \beta_t$. Thus, the only model required to train is a model which estimates the mean of the distribution μ_θ. This choice was made since it simplifies the formulation while empirically demonstrating the same results as the alternative, which would be approximating the variance.

In order to train the model so that $p_\theta(x_0)$ learns the data distribution $q(x_0)$, we can optimize the variational bound on negative log-likelihood as expressed in Equation 2.8,

$$\begin{aligned} \mathbb{E}[-\log p_\theta(x_0)] &\leq \mathbb{B}_q \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)} \right] \\ &= \mathbb{E}_q \left[-\log p(x_T) - \sum_{t \geq 1} \log \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_t \mid x_{t-1})} \right] \\ &= -\mathcal{L}_{\text{vlb}} \end{aligned} \tag{2.8}$$

where *vlb* stands for variational-lower-bound.

Ho *et al.* [32] found that instead of parameterizing $\mu_\theta(x_t, t)$ as a neural network, training a model $\varepsilon_\theta(x_t, t)$ which predicts ε, the noise that was added during a forward-step t, yielded

empirically better results.

$$\mu_\theta(x_t,t) = \tilde{\mu}_t\left(x_t, \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\cdot\varepsilon_\theta(x_t)\right)\right) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\cdot\varepsilon_\theta(x_0,t)\right) \tag{2.9}$$

Typically, ε_θ is represented and learned using a U-Net [85] architecture. The authors empirically found that the loss function worked best when simplified to the formulation demonstrated in equation 2.10,

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0,\varepsilon}\left[||\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t}\cdot x_0 + \sqrt{1-\bar{\alpha}_t}\cdot\varepsilon, t)||^2\right] \tag{2.10}$$

2.3.3.3 Variance Schedule

A key element which defines how samples are generated is the variance schedule, $\{\beta_t \in (0,1)\}_{t=1}^T$. It defines at which rate noise should be added along the Markov Chain. The schedule proposed by Ho *et al.* [32] increases linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ where $T = 1000$. Although simple, the authors considered it an effective approach. Yet, Nichol *et al.* [71] highlights how each diffusion step is not equally as important. In particular, the authors note that when $t \to 0$ there are a lot of unnecessary steps which, instead of improving the image, only introduce additional noise, which is especially noticeable in low-resolution images. To fix this, the authors propose using a cosine-based variance schedule (Equation 2.11) which empirically improved the model's performance.

$$\bar{\alpha}_t = \frac{f(t)}{f(0)},\ f(t) = \cos\left(\frac{t/T + s}{1+s}\cdot\frac{\pi}{2}\right)^2 \tag{2.11}$$

2.3.4 Denoising Diffusion Implicit Models

Song *et al.* [95] proposes a deterministic sampling process, Denoising Diffusion Implicit Models (DDIM), which is Non-Markovian, leading to faster sample generation.

Compared to DDPM, this method requires fewer steps to generate high-quality samples and is "consistent" since samples conditioned on the same latent variable share similar high-level features. Since these latents are considered semantically meaningful, it is also possible to interpolate two different latents in order to shift specific high-level features.

Based on previously proposed DDPM, a sample x_{t-1} can be generated from a sample x_t based on Equation 2.12,

$$x_{t-1} = \sqrt{\alpha_{t-1}}\underbrace{\left(\frac{x_t - \sqrt{1-\alpha_t}\varepsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}}\right)}_{\text{"predicted } x_0\text{"}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\varepsilon_\theta^{(t)}(x_t)}_{\text{"direction pointing to } x_t\text{"}} + \underbrace{\sigma_t\varepsilon_t}_{\text{random noise}} \tag{2.12}$$

For DDIM the authors define $\alpha_0 = 1$ and $\varepsilon_t \sim \mathcal{N}(0, \mathbf{I})$. When $\sigma_t = 0$, $\forall t \in \{0..T\}$, the process becomes deterministic, which is the key advancement proposed achieved by this model. Even though no longer Markovian, this model is still trained with the same objective as DDPM.

To further improve sampling time, the authors propose an *Accelerated Generation Process*. Since the denoising objective is not dependent on the forward process, it is possible to either propose a schedule with fewer T steps or, during the reverse process, skip steps to make the generation more efficient. These improvements require an analysis of trade-offs between efficiency and speed since the improvement of one tends to impact the other negatively.

2.4 Federated Learning

Training meaningful machine learning models typically requires a large amount of annotated data. Even though there are mechanisms to make training more data-efficient, such as Data Augmentation [77, 92] or Self-Supervised Learning [9, 10, 25], gathering the required data is still an expensive and time-consuming process. This makes sharing data between organizations and institutions to build better models collaboratively an appealing approach. Yet, this sharing brings up an onslaught of legal and privacy-related issues which make it difficult to apply in practice.

Additionally, in the medical field, there is an added challenge where individual hospitals cannot create meaningful models for rare medical conditions due to a lack of data. Thus, combining information from multiple data sources permits the creation of models which are both more generalized and unbiased than their counterparts [98].

Federated Learning (FL) [61] solves both issues by providing an approach to collaboratively train ML models without sharing data across institutions. Most FL frameworks for medical applications follow the structure shown in Figure 2.3, where each individual client trains an individual local model, whose information is then aggregated onto a global model located on a central server.

There have been multiple studies in the aforementioned domains for multiple applications such as COVID-19 detection through CT scans [19], chest radiography classification [44], brain tumour segmentation [90] amongst others. These works illustrate how the use of FL is viable in the medical field, leading to well-performing models.

Although classification tasks are most commonly studied, FL can also be applied to other tasks such as image generation using both GAN models [29, 57], diffusion models [42] and explanation retrieval [67].

In general, the global model is built by aggregating the parameters [96], which have been trained locally by the clients such as in Equation 2.13,

$$w_t^k = \sum_{k=1}^{K} \frac{n_k}{n} w_t^k \tag{2.13}$$

where there are K clients, n_k the number of datum belonging to client k and n the total number of data points, $n = \sum_k n_k$.

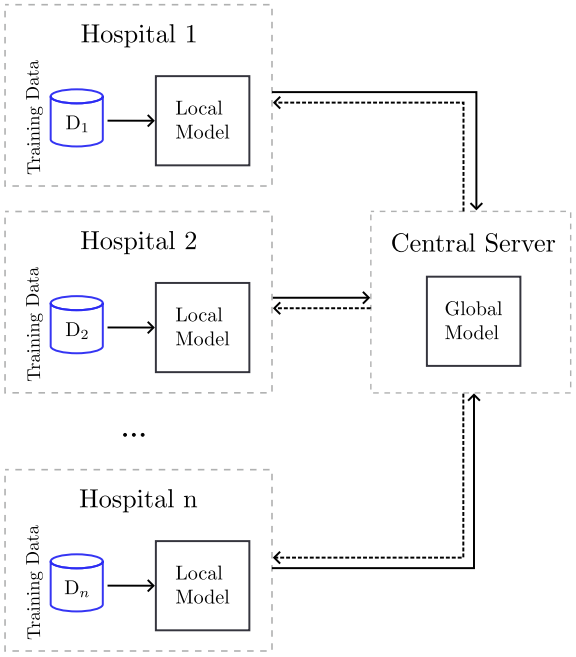


Figure 2.3: Federated Learning Framework.

This simplest approach FedSGD [61] requires each k client to take a step of gradient descent, after which the server will aggregate the weights. A drawback of this process is that the procedure is heavily serialized, relying on constant communication and being bottlenecked by the slowest client. The FedAvg [61] technique lessens this issue by permitting each client to perform multiple steps before broadcasting back to the central server. Multiple techniques have improved upon these original methods by promoting a more uniform distribution [56], improving the training stability [55] or leading to faster convergence [82].

Associated with FL is the concern of whether it is possible to extract the training data or information about it from the global model's weights and gradients. A common solution to these issues is the use of Differential Privacy mechanisms [1, 43, 106], which provides some guarantees that it will be unlikely to determine the original training data solely based on weights. In Section 3.5, we provide an in-depth analysis of attack vectors and potential risks which threaten privacy in both FL and non-FL settings.

2.5 Summary

In conclusion, the sharing of case-based explanations necessitates a preliminary step of visual anonymization to uphold privacy standards. Visual anonymization involves the obfuscation of identifying features in such explanations while preserving their realistic attributes, a task distinct from conventional methods like Differential Privacy. Notably, generative models emerge as an important tool for achieving this objective, with diffusion-based models garnering attention for

their recent strides in producing high-quality synthetic images when compared to more traditional approaches such as GAN and VAE.

By adopting diffusion-based models, such as DDPM or DDIM, which excel in capturing details while maintaining realism, the resultant anonymized case-based explanations not only safeguard patient privacy but also uphold the integrity of the medical data. This paradigm shift opens avenues for ethical information sharing, especially within the context of Federated Learning.

Integrating anonymized case-based explanations into Federated Learning frameworks holds promise for improving collaborative healthcare initiatives. The combination allows hospitals and healthcare institutions to share diverse disease cases without compromising patient privacy. This collaborative approach facilitates a collective understanding of complex medical scenarios, fostering advancements in diagnosis, treatment, and overall healthcare outcomes.

Moreover, this collaborative framework enables institutions with varying expertise and resources to collectively contribute to and benefit from a shared pool of anonymized medical cases. Consequently, the democratization of medical knowledge ensures that advancements and breakthroughs in healthcare are not confined to a select few but are accessible and beneficial to a broader spectrum of healthcare practitioners.

In summary, the convergence of visual anonymization through generative models and Federated Learning frameworks represents a transformative leap towards ethically sound, privacy-preserving information exchange in the domain of computer vision applied to healthcare. This paradigm not only facilitates collaborative learning among healthcare institutions but also upholds the sanctity of patient privacy, ultimately contributing to the advancement of medical science and the improvement of patient care on a global scale.

Chapter 3

Literature Review

3.1 Overview

In this Chapter we build upon the foundational knowledge displayed in the previous chapter by analyzing recent research on topics related to this work. With the goal of developing a system to generate anonymous case-based explanations there are several factors which will be taken into consideration:

- **Image quality** - the main objective of case-based explanations is to be reviewed by human practitioners. Therefore the explanations must have a high amount of detail and be realistic. Furthermore, the evaluation must be divided into two families of methods, quantitative and qualitative.
- **Utility** - an anonymized explanation should still be classified as the desired class by deep-learning-based methods. This is a necessity because an explanation with excellent image quality will be useless unless it represents the desired diagnosis.
- **Privacy** - the solution must be resilient to attacks that attempt to de-privatize the anonymous data. Ideally, there should the de-anonymization risk for approaches should be quantifiable.

Starting with Section 3.2, we present recent improvements and new use cases for Diffusion Models presented in the previous chapter. In Section 3.3 we present privacy-preserving methods which aim to visually anonymize images. In Section 3.4 we define what interpretability is and highlight key works which explore ways to achieve it, with a special focus on Case-based interpretability. Due to the importance of understanding the risks associated with anonymization methods in Section 3.5 we analyze different attack methods explored in Literature which aim to break anonymization targeting different aspects of the system. Finally, in Section 3.6 we will draw some closing remarks about the contents shown.

3.2 Diffusion Models

Multiple works build upon the foundation knowledge we analyze in the previous chapter. In this section, we will highlight the main avenues that have been explored, ranging from improving image quality to performing semantic manipulations in images using diffusion models.

3.2.1 Classifier-Guidance in Diffusion Models

As presented in Section 2.3.3, [32] proposes learning only the mean of the distribution and setting the variance to $\sigma_\theta(x_t,t) = \sigma^2 I$ in the reverse diffusion process, $\mathcal{N}(\mu_\theta, \sigma^2 I)$. More recently, [71] claims that, even though this choice does not affect sample quality, it does affect the log-likelihood. Therefore, the authors propose defining the variance as shown in Equation 3.1,

$$\Sigma_\theta(x_t,t) = \exp(v \log \beta_t + (1-v)\log \tilde{\beta}_t) \tag{3.1}$$

This defines a linear interpolation between β_t and $\tilde{\beta}_t$, where $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta$.

Since $\mathcal{L}_{\text{simple}}$ does not depend on $\Sigma_\theta(x_t,t)$ the original objective function must be redefined as:

$$\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{simple}} + \lambda \mathcal{L}_{\text{vlb}} \tag{3.2}$$

where $\mathcal{L}_{\text{vlb}}$ is the variational lowerbound as defined in Equation 3.3,

$$\mathcal{L}_{\text{vlb}} = \sum_{t=0}^{T} \mathcal{L}_t \tag{3.3}$$

$$\mathcal{L}_0 = -\log p_\theta(x_0 \mid x_1) \tag{3.4}$$

$$\mathcal{L}_{t-1} = D_{\text{KL}}(q(x_{t-1} \mid x_t, x_0) \mid\mid p_\theta(x_{t-1} \mid x_t)) \tag{3.5}$$

$$\mathcal{L}_T = D_{\text{KL}}(q(x_T \mid x_0) \mid\mid p(x_T)) \tag{3.6}$$

Building upon this foundation [15] proposes a classifier-guided sampling procedure for DDPM and DDIM alongside some empirically tested architecture improvements. This work filled the gap between GAN models, which have extensively been label-conditioned in different works in order to generate images conditionally. Given a classifier $p_\phi(y \mid x_t, t)$ trained on noisy images x_t for any timestep t, the gradients $\nabla_{x_t} \log p_\phi(y \mid x_t, t)$ guides the diffusion sampling process towards a class label y.

This is incorporated into the reverse process by shifting the mean of the distribution as defined in Equation 3.7.

$$x_{t-1} \sim \mathcal{N}(\mu + s\Sigma\nabla_{x_t} \log p_\phi(y \mid x_t, t), \Sigma) \tag{3.7}$$

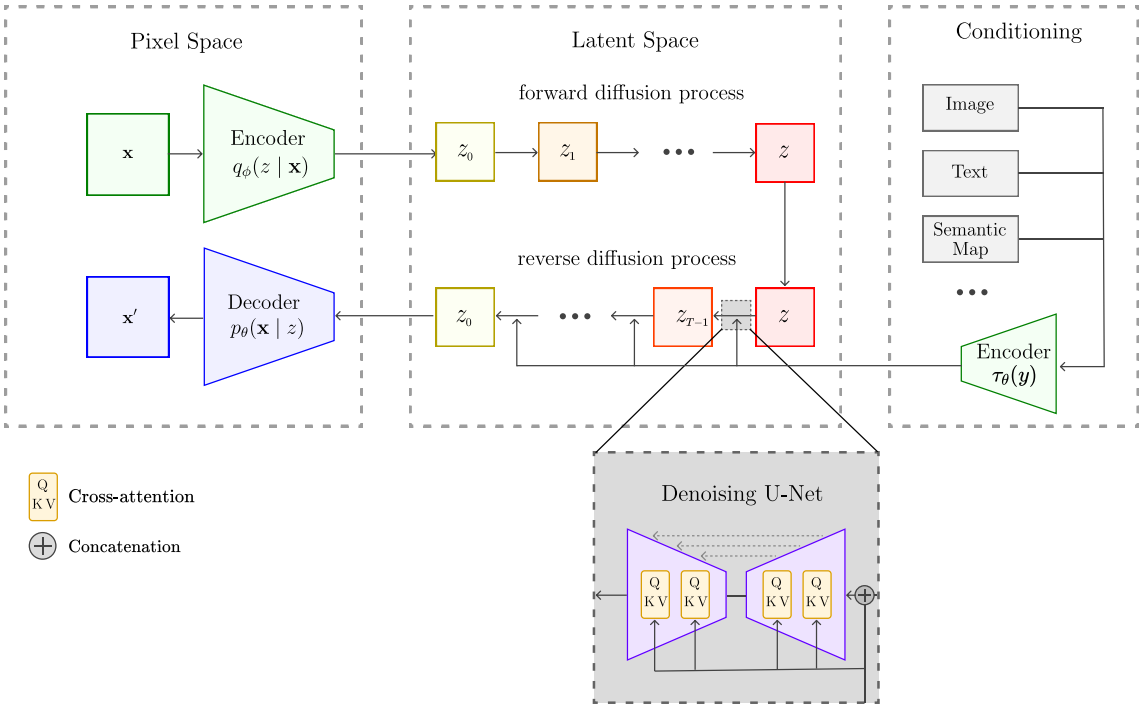


Figure 3.1: Latent Diffusion Model architecture. An image x is initially encoded into a latent z_0. This latent goes through the diffusion process, being noised into z_T and de-noised back to z_0, which can be decoded back into image space, creating a clean image x'. During the reverse diffusion process, the latent can be conditioned on a latent obtained by encoding any set of information.

An important downside of this process is that it assumes that a classifier p_ϕ is capable of correctly classifying noisy images, which, when $t \sim T$ are nearly pure Gaussian noise. This limitation is analyzed and worked around in works which build upon this foundation, such as the method proposed by Jeanneret *et al.* [37].

3.2.2 Latent Diffusion Models

As image size increases, so does the cost of both training and inference of DDPMs. Latent Diffusion Models (LDM) [84] employ a two-stage approach to dampen this issue. It consists of combining an autoencoder with a diffusion model, as illustrated in Figure 3.1, and is decomposed into the following set of steps:

1. Encode the original image into a latent variable
2. Execute the diffusion process using that latent variable as input. For every step of the de-noising process, the authors also employ a cross-attention mechanism [101] to condition the image generation.
3. Decode the latent variable back into pixel-space.

Steps 1 and 3 collectively form one stage, while Step 2 constitutes the other stage. The primary advantage of this method lies in executing the diffusion process on a low-dimensionality latent

rather than on a high-resolution image. This enables the generation of high-resolution images without the usual computational cost of applying the diffusion process to detailed images.

3.2.2.1 Conditioning

A key aspect of this approach is the **conditioning mechanism** to permit the modelling of conditional distributions following the form $p(z \mid y)$. This occurs in the auto-encoder employed in Step 2, which predicts noise following $\varepsilon_\theta(z_t, t, y)$, allowing for some control over the synthesis process through input y.

Since y can be in one of several modalities, it must first be encoded into an intermediate representation. The encoder τ_θ projects y to the representation $\tau_\theta(y) \in \mathbb{R}^{M \cdot d_\tau}$. This is then mapped to the intermediate U-Net layers via a Cross-Attention layer [101], which implements equation 3.8,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \tag{3.8}$$

where Q, K and V are defined by Equations 3.9:

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t),\ W_Q^{(i)} \in \mathbb{R}^{d \cdot d_\tau} \tag{3.9a}$$

$$K = W_K^{(i)} \cdot \tau_\theta(y),\ W_K^{(i)} \in \mathbb{R}^{d \cdot d_\tau} \tag{3.9b}$$

$$V = W_V^{(i)} \cdot \tau_\theta(y),\ W_V^{(i)} \in \mathbb{R}^{d \cdot d_\varepsilon^i} \tag{3.9c}$$

Where W_Q, W_K and W_V are learnable projection matrices. Based on this, the conditional LDM is obtained via Equation 3.10,

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\varepsilon(x), y, \varepsilon \sim \mathcal{N}(0,1), t}\left[|| \varepsilon - \varepsilon_\theta(z_t, t, \tau_\theta(y)) ||^2 \right] \tag{3.10}$$

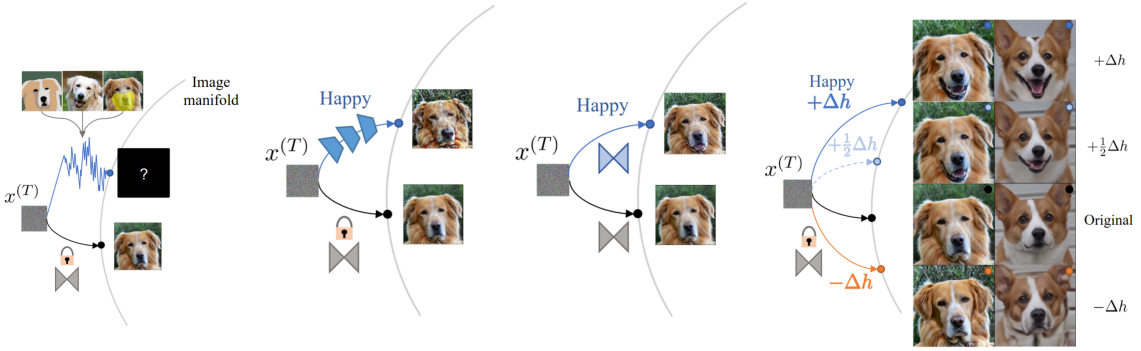
where τ_θ and ε_θ are both jointly optimized.

3.2.3 Medical Image Generation with Diffusion Models

The advent of diffusion models has led to their application in medical image generation tasks [46].

Cheff [103] Generates realistic chest radiographs by employing a novel architecture consisting of two main components: an LDM and a super-resolution network. Latent noise is de-noised and decoded by the latent diffusion model on an initial phase while optionally conditioned on information such as pathology labels or radiological reports. Afterwards, the output image is passed to the super-resolution network, inspired by [86], which will upscale and refine the image.

Brain Imaging Generation [78] This work generates 3D brain MRI images which outperform previous GAN-based methods. Similar to Cheff, the authors use an LDM, additionally conditioning the generation on covariables, such as brain structure volumes, age and sex.



(a) Image Guidance (b) Classifier guidance (c) DiffusionCLIP (d) Asyrp

Figure 3.2: Overview of manipulation approaches for diffusion models. Source: [54]

Medfusion [68] Based on LDM, Medfusion propose a generic method which is applied to different medical imaging domains. The main modification is the different number of channels in the VAE used to encode and decode latents. While the original LDM work used 4 channels, the authors found that using 8 channels led to a lesser number of visual artefacts.

In summary, the adoption of LDM over traditional approaches like DDPM and DDIM is evident, showcasing its effectiveness in advancing the capabilities of medical image generation models.

3.2.4 Semantic Latent Manipulation in Diffusion Models

One extensively studied aspect of Generative Adversarial Networks (GANs) pertains to their semantic latents. Specifically, manipulating these latent variables can yield meaningful semantic changes in the generated images [45]. However, applying similar manipulations to diffusion models proves to be more intricate due to the sequential nature of the de-noising procedure. Regardless of these difficulties, many authors have explored ways of bridging this gap in functionality. An overview of these methods is illustrated in Figure 3.2.

The simplest approach consists of using classifier guidance, as presented in the previous Subsection 3.2.1.

Diffusion Autoencoders Preechakul *et al.* [79] proposes a mechanism which allows, for each image, to obtain a "meaningful latent code" containing high-level semantic information. Firstly, a latent variable z_{sem} is obtained from a semantic encoder, which receives the original image as input $z_{\text{sem}} = \text{Enc}_\sigma(x_0)$. During the reverse diffusion process of a DDIM, the denoising U-Net is conditioned on this latent variable. Therefore each reverse step is formally defined as $p(x_{t-1} \mid x_t, z_{\text{sem}})$. Figure 3.3 shows an overview of this process.

The critical insight behind this approach is that an image can be represented by two components: the semantically rich vector z_{sem}, capable of representing the high-level details of the image, and a stochastic variable x_T containing finer-grained details.

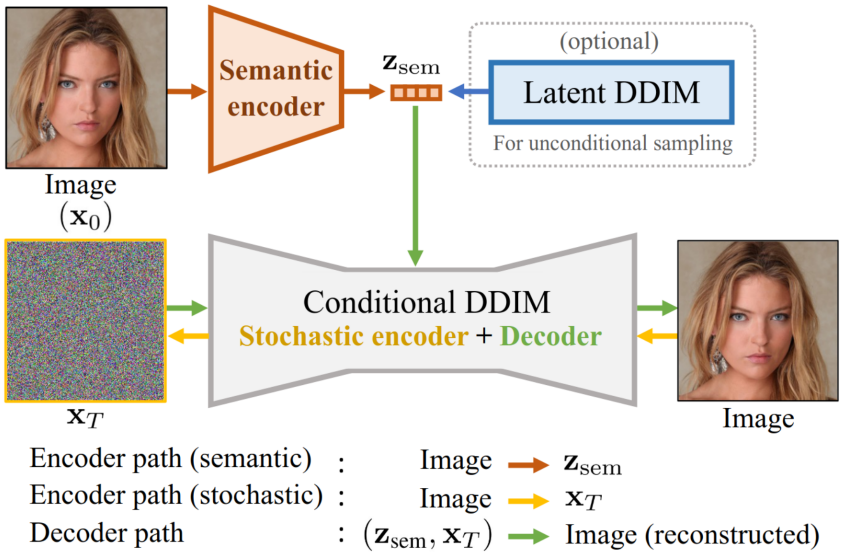


Figure 3.3: Architecture of a Diffusion Autoencoder. Source: [79]

The U-Net [85] is conditioned using adaptive group normalization layers (AdaGN). Then the AdaGN is conditioned on both t and z_{sem} as demonstrated in Equation 3.11:

$$\text{AdaGN}(h,t,z_{\text{sem}}) = z_s(t_s\text{GroupNorm}(h) + t_b) \tag{3.11}$$

Where $z_s \in \mathbb{R}^c = \text{Affine}(z_{\text{sem}})$ and $(t_s,t_b) \in \mathbb{R}^{2\times c} = \text{MLP}(\psi(t))$ is the output of a multilayer perceptron with a sinusoidal encoding function ψ.

Additionally, the authors have a process, which they denote as a stochastic encoder, to obtain a subcode x_T from an input image x_0. This is shown in Equation 3.12, where x_t is indirectly encouraged to encode information not already present in z_{sem}.

$$x_{t+1} = \sqrt{\alpha_{t+1}} f_\theta(x_t,t,z_{\text{sem}}) + \sqrt{1-\alpha_{t+1}}\varepsilon_\theta(x_t,t,z_{\text{sem}}) \tag{3.12}$$

This method allows for two types of image manipulations. Firstly, by varying the stochastic variable x_T for a constant semantic variable z_{sem}, the method can generate images with a similar structure compared to the original image while changing small details. Secondly, by directly changing the semantic variable z_{sem}, the authors showcase how specific characteristics of an image can be altered, for example, by modifying the smile of a person present in the picture. These modifications require moving the semantic variable in directions depending on the desired outcome. These directions are extracted from the weight vectors from classifiers trained to identify positive and negative instances of a specific attribute.

This also allows the interpolation of two images by interpolating their latent variables.

These models have in fact already been adapted to use-cases involving medical data. For instance Keicher *et al.* [48] models fracture grading as a continuous regression in the latent space of the DAE and shows its interpretability by visualizing different grades of a given vertebra. This

regression is calibrated to match the Genant scale, commonly used by medical professionals. The authors obtain an hyperplane $P = n \cdot w + b$, where n is the semantic direction, w an image and b a bias term, extracted from a binary classifier and manipulate the semantic latent z along the semantic direction n to shift the vertebra grade.

Asyrp Proposed by Kwon *et al.* [54], the asymmetric reverse process (Asyrp) discovers the semantic latent space, named *h-space*, in **frozen pre-trained** diffusion models. The authors claim that this semantic space has the following set of characteristics, which are desirable for semantic image manipulation tasks:

1. Homogeneity - The same Δh has the same effect across multiple samples.
2. Linearity - Scaling Δh linearly controls the magnitude of the change.
3. Robustness - The Δh should preserve the original image quality without degradation.
4. Consistency across timesteps - For different timesteps t, Δh_t is roughly consistent.

Recalling Equation 2.12 from the previous chapter, the authors first re-write and abbreviate that same equation as shown in Equation 3.13,

$$x_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{P}_t(\varepsilon_t^\theta(x_t)) + \mathbf{D}_t(\varepsilon_t^\theta(x_t)) + \sigma_t z_t \tag{3.13}$$

with $\mathbf{P}_t(\varepsilon_t^\theta(x_t))$ denoting the predicted x_0 and $\mathbf{D}_t(\varepsilon_t^\theta(x_t))$ the direction pointing towards x_t. The intuition behind this method lies in understanding that the objective is to modify the predicted image while maintaining the same "direction". For this purpose, the authors modify the equation so that a sample, shifted by Δh_t is obtained according to Equation 3.14.

$$x_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{P}_t(\varepsilon_t^\theta(x_t \mid \Delta h_t)) + \underbrace{\mathbf{D}_t(\varepsilon_t^\theta(x_t))}_{\text{direction pointing to } x_t} + \sigma_t z_t \tag{3.14}$$

where $\varepsilon_t^\theta(x_t \mid \Delta h_t)$ adds Δh_t to the original feature maps h_t.

3.3 Visual Anonymization

One key aspect of anonymizing case-based explanations is that many techniques typically used to assure data privacy are not applicable since the explanations must be exposed to humans to be useful. These techniques include Federated-Learning, Encryption or Differential Privacy, which, individually, do not fulfil the requirements for our problem. Taking that into consideration, throughout the years, different techniques have been proposed. They can be organised into two categories: the classic methods, which use simple techniques to obfuscate images and more recent strategies empowered by deep learning.

3.3.1 Classic Methods

Classic methods of performing visual anonymization include applying blurring techniques [22], such as Gaussian blur, to the images, reducing the detail level, or overlapping several images k (K-Same [69], K-Same-Select [26]) to guarantee K-Anonymity, ensuring that any attacker only has a $\frac{1}{k}$ probability of identifying a subject. With a large enough k, this guarantees anonymity. However, the definition of "large enough" is not adequately defined in laws and regulations.

The problem with these methods is that the visual quality degrades, making the images hard to use for tasks which require visual examination, such as case-based explanations, especially on applications where fine details are essential, such as MRI scans. This degradation can be visualized in Figure 3.4.

(a) K-Same

(b) K-Same-Select

Figure 3.4: Examples of applying K-Same and K-Same-Select. Source: [27]

3.3.2 Deep Learning Based Methods

The lack of classical approaches that do not hinder image quality motivated a shift toward approaches based on Deep Learning methods. This anonymization powered by the use of generative models can be separated into two major groups:

- **Post-model methods** - First, a sample is generated using a generic generative model. Afterwards, the sample is verified to check if it closely resembles a known identity and, based on that check, is either kept or removed from the dataset.
- **In-model methods** - In this approach, the generative model is aware of the goal of anonymization. Therefore, the generated samples should be anonymous.

Both approaches have their respective advantages and disadvantages. Post-model methods can be model-agnostic, making them easier to implement and apply to specific use cases. Meanwhile,

the other approach has the advantage of being more time-efficient because there will not be any time "wasted" generating images which might be scraped because they are not anonymous.

In the particular application scenario of FL, the method choice also has a significant impact on the framework's architecture. With post-model methods, all clients can synthesize their own anonymous dataset from which the central server will retrieve explanations. This puts both the computational expense and responsibility for the anonymization quality on each client. In contrast, with a post-hoc approach, the central server will be the one paying the generation toll and bearing the most responsibility in case of de-anonymization.

3.3.2.1 Diffusion-Based Methods

The use of diffusion models for visual anonymization is a relatively new field, having a limited number of works. Of these, many are not concerned with the anonymization quality, assuming that a completely synthetic image is, by definition, anonymous. These include the one proposed by Klempt *et al.* [52] which focuses on anonymizing pedestrian in car driving footage for self-driving applications or the one proposed by Luzi *et al.* [59] which provides a technique useful for many applications and attempts to extend it for anonymization purposes.

Another set of methods are the Differential Privacy-based methods [11, 17, 23]. Differential privacy provides stronger statistical guarantees that the training data cannot be reproduced when generating new samples. Yet, the current methods are unable to generate highly detailed images, which are extremely important for medical applications where details matter.

Packhäuser *et al.* [75] proposes a procedure to create anonymous chest radiographs using a post-model approach. This procedure demonstrated in Figure 3.5, consists of generating an entirely synthetic dataset, followed by the removal of synthetic images which closely resemble a patient's identity. This method consists of three main steps:

1. Train a generative model with the training data and use it to generate a large set of images. While any model can be selected for this purpose, the authors utilised an LDM for this step and compared it to a GAN baseline model.

2. For each generated radiograph, the method retrieves its top-1 image in terms of patient similarity from the training set. For this, the authors develop a network capable of converting the input radiographs into a meaningful embedding in such a way that embeddings from the same patient are closer together in the embedding space than ones from different patients. The network is a siamese neural network with two input branches, a ResNet-50 [30] with shared weights. It is trained using a contrastive loss to achieve a meaningful mapping from the high-dimensional image space to the low-dimensional latent space.

3. For each pair of original and generated images obtained in the previous step, a siamese network calculates the probability of both belonging to the same patient. If said value exceeds a threshold of $t = 0.5$, then that generated image is deemed to contain personal biometric patterns belonging to the patient and is removed from the dataset.

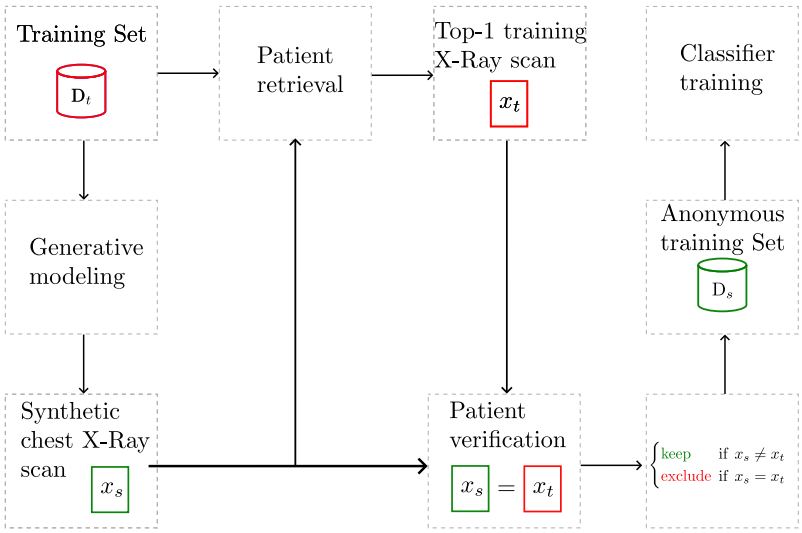


Figure 3.5: Architecture proposed by Packhäuser *et al.* [75] for the generation of anonymous chest radiographs. Source: [75]

3.4 Interpretability

Interpretability is an ill-defined concept, having multiple different definitions in the literature. Miller [62] describes it as "the degree to which a human can consistently predict the model's result". More recently, Doshi-Velez and Kim [18] define interpretability as the "ability to explain or to present in understandable terms to a human".

Interpretability refers to the capability of understanding the rationale behind decision-making processes. Deep Learning models commonly lack interpretability and are frequently characterized as "black boxes" owing to their large amount of nodes. This lack of interpretability poses a challenge, as these models can inadvertently magnify biases inherent in the initial dataset [28]. Additionally, these models can be susceptible to altered decisions by incorporating subtle variations that may not be discernible to humans [97], falling under the category of adversarial attacks. This can be especially damaging for medical applications.

Only a small set of existing models are deemed interpretable in literature. These include *decision trees*, *rule-based models* and *linear classifiers*. Unfortunately, for image classification tasks, these models are either unusable or lack performance compared to state-of-the-art approaches, which mostly focus on Deep Neural Network-based approaches. These models are classified under the broader category of **black-box models** since, by default, it is near-impossible to understand how they function.

Furthermore, interpretability approaches can be sub-divided into the following categories:

- **Pre-model approaches** - Focus on the understanding of underlying characteristics and distribution of the data.
- **In-model approaches** - These are built into the model. Compared to other approaches, these have more restrictions on their adoption, often requiring choosing and training specific

models from a limited set of options.

- **Post-model approaches** - These approaches provide a mechanism to explain decisions after receiving a classifier's prediction.

3.4.1 Case-Based Interpretability

Case-based explanations are a *post hoc* mechanism which justifies a model's decisions by retrieving cases or examples from the training data. This method is analogue to human reasoning, making them easy to interpret. These can be further refined into the following categories:

- **Typical examples** - A set of prototypical examples for a specific diagnostic will be retrieved. This is akin to consulting a reference document which documents different diagnoses.
- **Similar examples** - For this approach, a set of similar images with the same diagnosis will be retrieved. This is helpful since the same disease may manifest in many visually distinct ways.
- **Counterfactual examples** - A counterfactual explanation demonstrates what changes must be made to an image to alter a classifier's diagnosis. This approach can also show iterative changes, highlighting what aspects of the image are being altered, making them more easily interpretable.
- **Semi-factual examples** - This approach attempts to make the largest possible changes to an image while maintaining the same classifier prediction. This moves the image closer to the decision boundary, in a manner similar to counter-factual explanations.

3.4.1.1 Explanation Retrieval

While counterfactual explanations can be generated based on a single reference image, similar examples need to be retrieved from existing datasets. Logically, the quality of the explanations can be highly dependent on the method with which these are retrieved.

Typical image retrieval systems only need to rely on simple similarity metrics, such as the Euclidean distance between images, since the target is to obtain an image that, as a whole, looks similar to a reference input image.

Yet, this approach is not generally desirable in medical imaging data since the details that differentiate different diagnoses are typically spatially small, making them overlooked by distance metrics that value the entire image equally. This stimulated the need to develop new retrieval systems specifically designed for specific tasks.

Most approaches [14, 74] revolve around using image labels under a supervised setting. Commonly, the similarity measure is obtained through a disease-specific classifier. IG-CBIR (Interpretability-Guided Content-Based Medical Image Retrieval) [93] is a retrieval method that tries to shift the retrieval system focus to the image sections, which are most important for typical classification

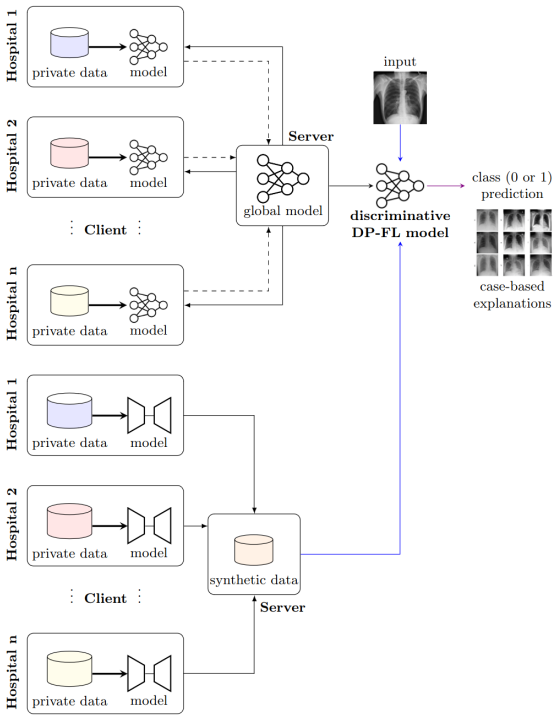


Figure 3.6: Pipeline for the DP-FL model. Source: [67]

tasks. To do so, for an input image, the method will first process it through a classifier network in order to retrieve its Deep Taylor saliency map [63]. In the second stage, this map is then processed through another disease classifier from which the feature vector present in the penultimate model layer will be extracted. Finally, images are compared and retrieved based on the Euclidean distance of their features. The authors concluded that this approach led to an improved retrieval quality compared to previously existing approaches.

Explanation retrieval has also been applied under FL setting by Laura *et al.* [67], with a novel model, DP-FL (Figure 3.6), which is based on the FedAvg approach. This work studies the use of case-based explanations to explain the presence or absence of Pleural Effusion which can be diagnosed based on chest X-ray images. In this work, the DP-FL model, which is trained collaboratively between hospitals, retrieves images from a synthetic dataset, which has also been created by multiple institutions. The key limitation of this approach is the lack of anonymization, which may lead to the leakage of private and sensitive medical data.

3.4.1.2 Anonymous Case-Based Explanations

Montenegro *et al.* [64] identify several gaps in existing privacy-preserving methods. In addition to the previously discussed classic methods, the authors critically analyze the PPRL-VGAN model [8], which anonymizes images through the use of a replacement identity. This approach has the shortcoming of exposing the replacement's identity.

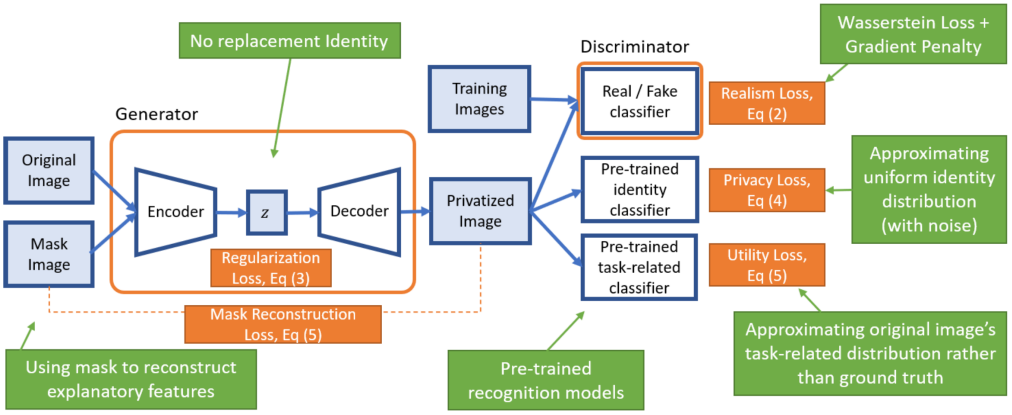


Figure 3.7: Overview of the privacy-preserving model proposed by Montenegro *et al.* [65] Source: [65]

The authors improve upon PPRL-VGAN by proposing a privacy-preserving model composed of three modules [65]: a privacy module, an explanatory module and a generative module. This architecture is shown in Figure 3.7.

The generative module, responsible for generating realistic images, employs a GAN, which uses a VAE as its Generator. The discriminator, D, is trained using a Wasserstein loss function [2] with an additional gradient penalty, demonstrated in Equation 3.15.

$$\mathcal{L}_D = \underbrace{\mathbb{E}_{I\sim p_d(I)}[D(G(I)) - D(I)]}_{\text{Wasserstein loss}} + \underbrace{\mathbb{E}_{\hat{x}\sim p_{\hat{x}}}[\lambda(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2]}_{\text{gradient penalty}} \tag{3.15}$$

where I is an image.

The generator comprises two terms: a realness term, in Equation 3.16, and a regularization term in the VAE, shown in Equation 3.17.

$$\mathcal{L}_{\text{realness}} = \mathbb{E}_{I\sim p_d(I)}[-D(G(I))] \tag{3.16}$$

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{I\sim p_d(I)}\left[D_{\text{KL}}\left(q\big(f(I)|I\big) \;||\; p\big(f(I)\big)\right)\right] \tag{3.17}$$

Two interchangeable **privacy modules** have been proposed, each with unique advantages and disadvantages.

The first employs promotes a uniform distribution in a pre-trained multi-class identity recognition network D_{id}. This means that, for any generated explanation, any patient identity is equally probable, providing similar privacy guarantees compared to the classic K-Same [69] methods. These considerations are taken into consideration in the loss term represented in Equation 3.18,

$$\mathcal{L}_{\text{privacy}} = \mathbb{E}_{I\sim p_d(I)}[-D_{id}(G(I))\log(U)] \tag{3.18}$$

where U represents a uniform distribution with noise.

The second approach employs a Siamese identity recognition network instead of the multiclass

classification method. This network determines whether two input images belong to the same patient or not, and is trained using the Contrastive Loss shown in Equation 3.19

$$\text{ContrastiveLoss} = \left(Y \times \text{ED}^2 + (1-Y) \times [\max(0, m - \text{ED})]^2\right) \cdot \frac{1}{2} \tag{3.19}$$

where ED is the Euclidean distance between the image pair embeddings and Y a label for the image pair, 1 if they share the patient identity and 0 otherwise.

A privatized image should have a large identity distance from every patient present in the dataset. This is taken into consideration by Equation 3.20.

$$\begin{aligned}\mathcal{L}_{\text{privacy}} = \mathbb{E}_{(I,N)\sim p_d(I,N)} \Bigg[& \lambda_2 \Big[\max(0, m - \text{ED}(I, G(I)))\Big]^2 \\ & + \lambda_6 \frac{1}{N} \sum_{i=0}^{N} \Big(\max(0, m - \text{ED}(G(I), I_N))\Big)^2 \Bigg]\end{aligned} \tag{3.20}$$

The **explanatory module** applies a task-related classifier D_{exp} to obtain Deep Taylor saliency maps [63], M. Additionally, this module also approximates the classification score of the privatized image to the original. These considerations are depicted in Equation 3.21.

$$\mathcal{L}_{\text{exp}} = \mathbb{E}_{(I,M)\sim p_d(I,M)} \left[\lambda_3 \underbrace{D_{\text{exp}}(I) \log\Big(D_{\text{exp}}(G(I))\Big)}_{\text{approximate classification score}} + \lambda_4 \underbrace{\Big(I \times M - G(I) \times M\Big)^2}_{\text{reconstruct saliency map}} \right] \tag{3.21}$$

The final loss term for the generator is shown in Equation 3.22,

$$\mathcal{L}_G = \lambda_1 \mathcal{L}_{\text{realness}} + \lambda_2 \mathcal{L}_{\text{privacy}} + \mathcal{L}_{\text{exp}} + \lambda_5 \mathcal{L}_{\text{reg}} \tag{3.22}$$

where λ_k is a weight associated with each term.

The major downside of this approach is the limited visual quality of the explanations it produces. Ideally, these should be as realistic as possible while maintaining anonymity.

3.5 De-Anonymization Methods

De-anonymization consists in uncovering the real identity of an individual based on the data which was previously anonymized. As explained in Chapter 2, within the GDPR [12] regulation, this can usually be attributed to singling out an individual in the dataset, even if its real identity is still unknown.

To assess the anonymization quality attained by the proposed solution, it is important to undergo a paradigm shift. Rather than analyzing the issue from the defender's standpoint, examining

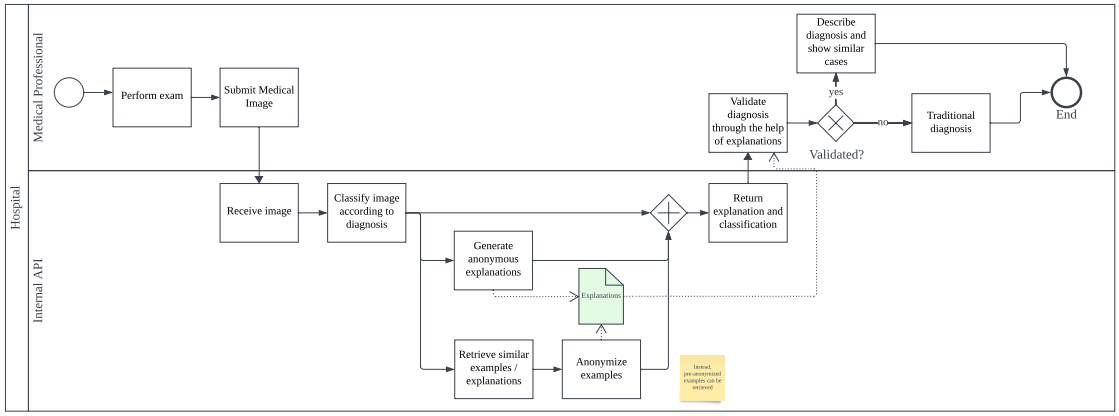


Figure 3.8: BPMN diagram [73] of the proposed workflow for a practical application of the method under study.

potential attackers' perspectives is required. This analysis must take into consideration different levels of access to data:

- **Explanation Access** - The attacker can access a set of anonymized explanations, and the attack vectors can be focused on de-anonymizing individual explanations. This is the most common access method, given a regular workflow, shown in Figure 3.8.
- **Black-box Access** - The attacker has access to a server API, which is the equivalent of having access to both the input and output of the model. This is also known as a black-box [91].
- **White-box Access** - Compared to the previous item, this considers access to the model weights themselves. This is a valid concern, especially when employing a Federated-Learning framework where one or more of the multiple centres involved in the training process is malicious.

We analyze different threat models under which our methodology may be at risk, studying what access level they require to be executed. The following threats are the ones considered most relevant.

Membership Inference Membership Inference Attacks (MIA) are commonly studied attacks in literature [91] and consist of verifying if a given data point was present in the training data of a model, in other words, inferring whether that datum was a member of the dataset. In the medical domain, the awareness of a patient's presence within the dataset could enable entities such as insurance companies to assess the likelihood of the patient being afflicted with a particular ailment.

Formally, considering $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, a training dataset composed of paired sets of images (x_i) and their respective labels (y_i). Given a model f we want to discover whether a new sample $k \in D$.

Duan *et al.* [20] analyses existing MIA which have been proposed for GAN models and studies their adaptability to diffusion models. Furthermore, the authors propose a novel approach to perform a MIA under a black-box setting. Expanding upon this earlier work, Kong *et al.* [53] propose an attack target at DDIM by predicting the denoising trajectory of a sample, which is only possible due to the deterministic nature of the models. Matsumoto *et al.* [60] explores both white-box and black-box settings and demonstrates that, for both, the membership inference accuracy decreases with the increase of the number of training samples. Hu *et al.* [33] proposes two different approaches: first, a loss-based attack targeted at models in a white-box setting and, secondly, an attack where the attackers have access to the likelihood values of samples from the diffusion model.

Data Extraction A privacy concern in Generative Models is the question of whether or not they can generate new samples which are identical to the training data used, therefore compromising the identity of the original patients. Considering a training dataset $D = \{x_1, x_2, ..., x_n\}$ this family of attacks focus on extracting an image $\hat{x}$ which fulfils the condition $\exists\, x \in D \;:\; \hat{x} \approx x$. This can be done either under a black-box or white-box setting and, in fact, Carlini *et al.* [7] study both approaches. These attacks are commonly combined with MIA to verify that, in fact, the generated sample $\hat{x}$ is similar to an existing training sample. It is an exemplary method of studying whether a generative model is capable of leaking the data used to train it.

Data Reconstruction Considering a subset of a training image, $k \subset x, x \in D$, this attack aims to reconstruct the original image x based solely on this subset k. This is a weaker formulation of the previous attack and is also applied to diffusion models by Carlini *et al.* [7].

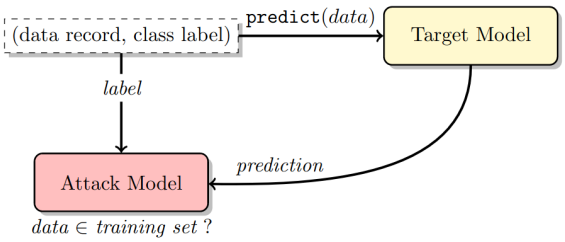


Figure 3.9: Diagram of membership inference attack in a black-box setting. Source: [91]

Multiclass Identity Recognition Network This type of network would receive an image and, based on its features, determine which patient (or class) the image belongs to. This requires knowing *a priori* a set of patients from which to predict and having enough data to train the network. These conditions make it not as practical as other approaches as a means of attacking. Yet, it is used in literature to analyze privacy-preserving models, making it, at the very least, an interesting benchmark.

Formally, considering an image x, the model f will predict the likelihood of the image belonging to patient k from a set of patients $K = \{k_0, k_1, ..., k_n\}$ Formally, a identity $\hat{y}$ will be predicted

Table 3.1: Overview of attack methods proposed in Literature.

Method	Paper	Year	Explanation Access	Black-Box Setting	White-Box Setting
Membership Inference	Duan *et al.* [20]	2023	-	-	✓
	Kong *et al.* [53]	2023	-	-	✓
	Matsumoto *et al.* [60]	2023	-	✓	✓
	Hu *et al.* [33]	2023	-	-	✓
Extract Data	Carlini *et al.* [7]	2023	-	✓	✓
Reconstruct Data	Carlini *et al.* [7]	2023	-	✓	✓
Multiclass Identity Recognition	Montenegro *et al.* [66]	2022	✓	-	-
Siamese Identity Recognition	Montenegro *et al.* [66]	2022	✓	-	-
	Packhäuser *et al.* [76]	2022	✓	-	-

based on expression 3.23,

$$\hat{y} = \underset{k \in K}{\operatorname{argmax}}\, f(x) \tag{3.23}$$

Siamese Identity Recognition Network Another approach involves training a Siamese network, a network which takes as inputs two images and predicts whether they belong to the same patient or not. This network has two branches, typically with shared weights and trained with a contrastive loss.

Formally, considering two images x_0 and x_1, the model f will predict $\hat{y} = f(x_0, x_1)$, whether they belong to the same patient or not.

This approach has the advantage over the multiclass classification approach by being more easily extensible to patients which were not contained in the original training data, which is the most common scenario. It is common because we can consider that an attacker can access any dataset except the one from the specific centre it is trying to target.

3.6 Summary

In conclusion, this chapter has delved into important considerations for the development of a system aimed at generating anonymous case-based explanations in the context of Computer Vision. The analysis began with a focus on image generation, emphasizing the necessity for detailed and realistic explanations, which is commonly achieved through the use of LDM. We also explore the commonly used approaches to perform semantic manipulations in diffusion models which is a task closely related to counterfactual explanations.

In terms of visual anonymization, we overviewed some common approaches to anonymize images used in literature and then narrowed it down to the under-explored area of anonymization using diffusion models.

We also underscore the importance of interpretability in DNN with a great focus on case-based explanations, providing examples of methods for obtaining both counter-factual explanations and similar/typical example explanations.

However, the chapter also underscored the significance of acknowledging and addressing potential risks associated with anonymization methods. Section 2.5 critically examined different attack methods documented in the literature, each targeting distinct aspects of the system.

Finally, we close this chapter by analyzing methods which have been used to re-identify or compromise the privacy of patients present in the medical training data used by various models.

Chapter 4

Preliminary Data Choices

4.1 Overview

The problem we are tackling and the solutions to solve it, which we propose in the following chapters, are task-agnostic. That is, the method should be applicable to any two-dimensional medical imaging classification task. For this reason, for our experiments, we elected to focus on chest X-ray classification since it is one of the most common types of medical imaging tasks and due to access to certified radiologists to quantitatively and qualitatively evaluate the quality of the generated images. This evaluation is an important step since the solution must be approved and aligned with the interests of its end users.

Additionally, there is an abundance of large chest X-ray datasets that can be used to perform experiments. A comparison of five of the most used datasets for these tasks is shown in Table 4.1. From this set, the only dataset that is unusable is VinDr-CXR [70] since it does not aggregate images by patient, therefore making it impossible to evaluate re-identification risks properly. To simulate a FL setting, we employ multiple datasets from the aforementioned list to emulate different hospitals.

4.2 Datasets

We focus this work on 3 specific datasets: MIMIC-CXR-JPG [41], BRAX [83] and CheXpert [35]. All three datasets have labels obtained using the CheXpert labelling tool, which guarantees label

Table 4.1: Chest X-ray dataset comparison.

Dataset	Image Count	Patient Count	File Format	Size (GB)
CheXPert [35]	224,316	65,240	Images	471.12
BRAX [83]	40,967	19,351	Image/DICOM	432.1
MIMIC-CXR-JPG [41]	377,110	65,379	Image	557.6
NIH ChestXRay [102]	112,120	30,805	Image	45.08
VinDr-CXR [70]	18,000	-	DICOM	205.96

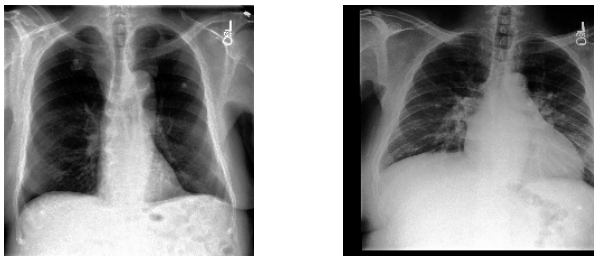
uniformity and makes combining them easier. The CheXpert labelling tool automatically extracts observations for 14 different conditions from free-text radiology reports, labelling them as blank (if unmentioned), -1 (if uncertain), 0 (if negative), and 1 (if positive). For simplicity, we only use images clearly identified as 0 or 1 (U-Ignore) for all experiments. While the authors of the dataset report that their best-performing model for Cardiomegaly classification uses a multiclass approach (U-MultiClass) in which uncertain labels are their own class, for simplicity for all the experiments, we only use images whose label is clearly identified, either 0 or 1 (U-Ignore). Each dataset has its own distinct characteristics:

- **MIMIC-CXR-JPG** [41] - It is a large-scale Chest-Xray dataset derived from the MIMIC-CXR [40] dataset by obtaining JPG images from the original DICOM files and structured labels from free-text reports.
- **BRAX** [83] - The Brasilian labelled chest x-ray dataset (BRAX) is an automatically labelled dataset containing a large number of studies performed in a general Brazilian hospital. All images have been verified by radiologists, and the labels were extracted from free-text radiology reports, originally written in Brazilian Portuguese, using NLP techniques.
- **CheXpert** [35] - CheXpert is a large dataset composed of radiology studies collected from Stanford Hospital from October 2002 to July 2017. The test set used by CheXpert was manually labelled by 5 radiologists. This dataset is notable for its size and the high quality of its labels, which have been validated by multiple experts.

From these three, we decided to use MIMIC-CXR-JPG to perform most experiments in a centralized setting and include the remaining datasets to emulate multiple hospitals for experiments which employ Federated Learning. The combination of these datasets provides an accurate simulation of a real federated learning scenario since they span multiple geographical regions and sensor types.

4.3 Task choice

We elected to focus on the Cardiomegaly disease since it is common, visually identifiable and available in all the aforementioned datasets. Cardiomegaly is a common heart disease characterized by an enlarged heart. The most common analysis method is through the analysis of posteroanterior (PA) chest X-rays since the anteroposterior (AP) view tends to magnify the heart [80] hindering the diagnosis. More specifically, a patient is diagnosed with Cardiomegaly if its PA chest X-ray shows a cardiac silhouette that is greater or equal to 50% of the diameter of the chest. Figure 4.1 showcases the visual difference between a healthy X-ray and one which presents a Cardiomegaly diagnosis.



(a) No Cardiomegaly. (b) Cardiomegaly.

Figure 4.1: Comparison between a healthy X-ray and one that presents Cardiomegaly. The key difference that a radiologist must look out for is the enlarged heart area. Both images were obtained from the MIMIC-CXR-JPG dataset.

4.4 Summary

In this chapter, we discussed the selection process for datasets and the specific focus of our experiments. We selected chest X-ray classification as our primary task due to its prevalence and the availability of large, high-quality datasets. The datasets chosen – MIMIC-CXR-JPG, BRAX, and CheXpert – provide a solid foundation for both centralized and federated learning experiments. By focusing on Cardiomegaly, we ensure that our experiments are relevant and aligned with real-world clinical needs. The following chapters will delve into the methodology and experimental results, demonstrating the efficacy and privacy-preserving nature of our proposed solution.

Chapter 5

In-model Anonymization

5.1 Overview

In-model anonymization is a family of methods for anonymous image generation, where the generative model is tasked with creating realistic images while being distinct from the training samples. This method can then be used to create an anonymous dataset, which will serve as a case-based explanation catalogue, as shown in Figure 5.1. This approach is appealing since it is more computationally efficient than the alternative post-model anonymization approach, which will be presented in Chapter 6. In literature, the task of generating anonymous images has been largely unexplored. Therefore, in this chapter we propose three distinct methods which attempt to solve this task.

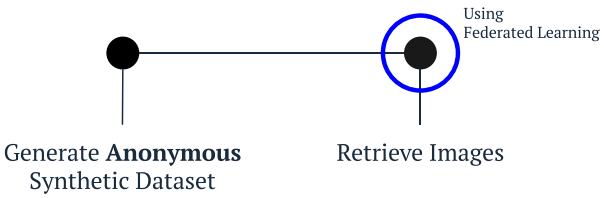


Figure 5.1: In-model approach: generate anonymous synthetic images and retrieve them.

5.2 Differentially Private Diffusion Model

Differential privacy, a robust privacy-preserving technique, has been effectively implemented in diffusion models through the DPDM [17] model. While this method offers strict privacy guarantees for generated images, it is often associated with a trade-off in image quality. Although this limitation is known, we intend to explore the viability of using this method to generate case-based explanations.

To formally define differential privacy [39], consider two datasets D_1 and D_2, that differ by only one element. These datasets are called *neighbouring datasets*. An algorithm $\mathcal{M}$ satisfies

(ε, δ) – differential-privacy if and only if for all neighbouring datasets and for all subset S of $\mathcal{M}$'s range, the condition shown in Equation 5.1 holds:

$$\Pr[\mathcal{M}(D_1) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D_2) \in S] + \delta \tag{5.1}$$

Both a lower privacy budget,ε, and a lower δ correspond to stronger privacy, typically at the cost of performance. The number δ can be interpreted as "the probability that a mechanism's output varies by more than a factor of e^{ε} when applied to a dataset and any of its neighbors". Typical values for ε are usually between 0.1 and 10 while δ can be determined using the heuristic $\delta = \frac{1}{N}$ where N is the number of samples in the dataset.

5.2.1 Experiments

A key limitation of the DPDM methods is that, as the image resolution of the output images increases, so does the expected training time due to increased complexity. This limitation occurs since networks which generate higher-resolution images require more parameters, making them problematic to train under differential privacy due to the amount of noise introduced, which makes it difficult for the model to converge. Therefore, we limit the resolution of each image to 64×64 pixels, the maximum resolution explored by the authors of the original work. For this experiment, we employ the MIMIC-CXR-JPG dataset.

We train the model, based on the UNet architecture [85], for 300 epochs using the Adam optimizer with a learning rate of $3e^{-4}$ and no weight decay. We chose to use ($\varepsilon = 10$, $\delta = 1e^{-6}$) as the privacy settings for this experiment. While using $0 \leq \varepsilon \leq 1$ is ideal to assure patient privacy, we elected to use $\varepsilon = 10$ to obtain baseline results that help preserve data privacy while trying to obtain acceptable image quality. We use the MIMIC-CXR-JPG to train this model.

Another limitation of DPDM is that it considers per-image privacy guarantees. Yet, in all datasets, each patient may contain multiple images, turning this into a Group Privacy problem. Instead of guarantees for (ε, δ)-DP instead for a patient with k images we have ($k\varepsilon$, $ke^{(k-1)\varepsilon}\delta$)-DP. Considering the MIMIC-CXR-JPG dataset, the upper bound for k would be 158 since that is the maximum number of studies available for a single patient. One way of improving privacy concerns and decreasing k is by limiting the amount of training data to N studies per patient.

We generate example samples using DDIM with $T = 50$ steps during the inference process for qualitative examination. While we can generate anonymous images, it is difficult to employ them as case-based explanations due to their low quality. They simultaneously have low resolution and present large visual artefacts that hinder their analysis, as shown in Figure 5.2. For this reason, we are required to propose different anonymisation methods that can achieve more realistic images.

5.3 Decoder-level Anonymization

Decoder-level anonymization employs an LDM and removes the generated images' personally identifiable information during the decoding step, in which a decoder D maps a latent variable

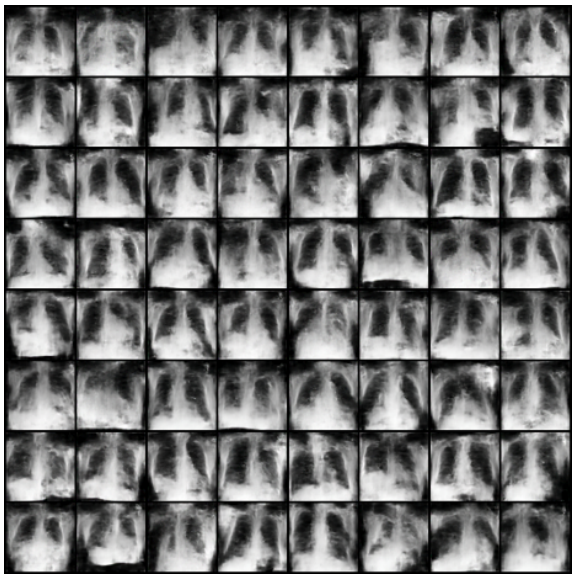


Figure 5.2: Samples generated using the DPDM model.

back into the image space. Figure 5.3 shows an overview of this method.

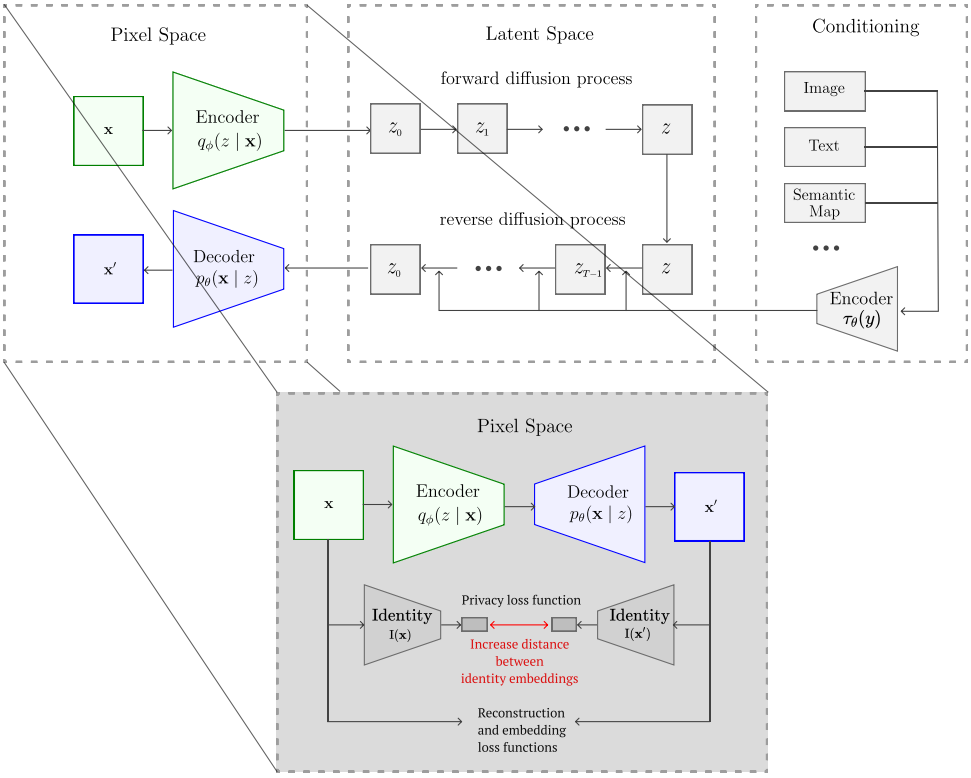


Figure 5.3: Overview of the training process for the decoder-level anonymization method. The VAE is trained so that the input and output images have a large identity-wise distance. The identity embeddings of the images are obtained using a patient retrieval network, I.

5.3.1 Method

We modify the VAE training procedure proposed by Müller-Franzes *et al.* [68] by adding a new loss function that promotes anonymity. This loss function, $\mathcal{L}_{\text{privacy}}$ is similar to the one proposed by Montenegro *et al.* [66]. Formally, the VAE used in the LDM is trained according to the loss function shown in Equation 5.2,

$$\mathcal{L}_{\text{VAE}} = \underbrace{\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{LPIPS}} + \mathcal{L}_{\text{SSIM}}}_{\mathcal{L}_{\text{reconstruction}}} + \mathcal{L}_{\text{embedding}} \tag{5.2}$$

Which consists of a reconstruction loss based on Mean Squared Error (MSE), the Learned Perceptual Image Patch Similarity (LPIPS) and the Structural similarity index measure (SSIM) and an embedding loss which optimizes the KL divergence between the embedding space and white Gaussian noise.

Recalling the privacy loss proposed by Montenegro *et al.* [66], to promote the privacy of patients, we use the loss function proposed in Equation 5.3.

$$\begin{aligned}\mathcal{L}_{\text{privacy}} = \mathbb{E}_{(x,N)\sim p_d(x,N)}\Bigg[&\lambda_1\Big[\max(0, m - \text{ID}(x,\hat{x}))\Big]^2 \\ &+ \lambda_2 \frac{1}{N}\sum_{i=0}^{N}\Big(\max(0, m - \text{ID}(\hat{x}_N, x_N))\Big)^2\Bigg]\end{aligned} \tag{5.3}$$

where m is a margin parameter, $\text{ID}(x,y)$ is an identity distance function which computes the distance between identity embeddings of two images. In practice, we want a generated image, $\hat{x}$, to have a high identity distance from the original image x. Additionally, this generated image should also be distant identity-wise from the remaining training images, $\{x_N : \forall N \in D\}$.

Finally, the final objective of our privacy-preserving VAE is given by Equation 5.4.

$$\mathcal{L}_{\text{VAE}} = \underbrace{\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{LPIPS}} + \mathcal{L}_{\text{SSIM}}}_{\mathcal{L}_{\text{reconstruction}}} + \mathcal{L}_{\text{embedding}} + \mathcal{L}_{\text{privacy}} \tag{5.4}$$

With this procedure, we aim to remove the personally identifiable details of images during the decoding step of the generation process. Effectively, this approach has two competing goals: recreate the input image as accurately as possible and increase the identity distance of the input and output as much as possible.

As a stepping stone towards the defined goal, we first develop a baseline approach which uses a simplified privacy loss function. Equation 5.5 defines this loss function, which attempts to increase the identity distance between the original and the reconstructed image.

$$\mathcal{L}_{\text{simplified\_privacy}} = \mathbb{E}_{(x,N)\sim p_d(x,N)}\Big[\lambda_1\Big[\max(0, m - \text{ID}(x,\hat{x}))\Big]^2\Big] \tag{5.5}$$

Table 5.1: Results of training VAE with privacy loss function. Reported on the validation set.

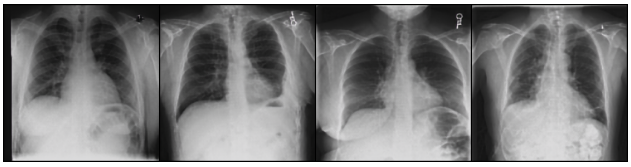
λ_1	m	SSIM (↑)	MSE (↓)
0.001	5.0	0.91328	0.00140
0.001	10.0	0.91572	0.00141
0.01	5.0	0.91222	0.00136
0.01	10.0	0.91536	0.00143
0.1	5.0	0.91110	0.00146
0.1	10.0	0.91439	0.00136
1.0	5.0	0.91515	0.00135
1.0	10.0	0.91462	0.00135
5.0	5.0	0.91163	0.00145
5.0	10.0	0.91420	0.00138

5.3.2 Experiments

We start by modifying the Medfusion architecture by adding the newly proposed $\mathcal{L}_{\text{simplified\_privacy}}$ loss function. This function includes two hyperparameters which must be defined, λ_1 and m, which are a weight constant and a margin value, respectively. To determine which values work best, we perform a grid search and train the architecture for $(\lambda_1, m) \in \{0.001, 0.01, 0.1, 1, 5\} \times \{5, 10\}$. While the order of magnitude of m could be determined by analyzing identity embeddings obtained from the training set, it was harder to determine which order of magnitude worked best for the weight, which explains why the search space is larger. We also define the $\text{ID}(x, y)$ as being the Euclidean distance between the embeddings $I(x)$ and $I(y)$ obtained using a retrieval network, I, proposed by Packhäuser *et al.* [75] trained for 30 epochs in the first phase, during which the model backbone is frozen, and 50 epochs during the second phase where all parameters are unfrozen. We use the learning rate suggested by the paper's authors, 0.158489, with weight decay of $1e^{-5}$ and batch size 32. This retrieval mechanism is further analyzed in Chapter 6. Finally, our VAE model is tasked with recreating 512×512 pixel images and was optimized using Adam [49] using a learning rate of $1e^{-4}$.

The result of this method is, effectively, an adversarial attack which introduces noise to the output image in the form of blur without structurally altering the input image. While the change of hyperparameters impacts the reconstructed images, this impact is limited to a change in blurring strength; this limited impact is also corroborated by the results shown in Table 5.1. Even though this approach may lead to images capable of fooling a patient retrieval model, they are not ideal as case-based explanations for two reasons: the images may still be de-anonymized by a different model, and their visual quality is not ideal for visual examination.

(a) Input images.



(b) Output images.

Figure 5.4: Comparison between input and output images using VAE trained using the loss function proposed in Equation 5.4. The output images are structurally identical to the originals, varying only in subtle colouration of blur changes, which adversarially confuses the retrieval model.

5.4 Classifier-free Identity Guidance for Privacy-Preserving Image Generation

To avoid the adversarial nature of the previously proposed decoder-level approach, we instead turn our focus towards the diffusion process of the LDM. In Chapter 3 we discussed multiple approaches which steer the diffusion model in such a way that the generated images exhibit desired semantic characteristics. During this section, we experimented with modifying one such mechanism, classifier-free guidance, for the purpose of image anonymization.

5.4.1 Method

To build up towards creating an identity-guided classifier-free guidance mechanism, we first overview the original formalization of said technique and expand it to create a simpler task.

5.4.1.1 Classifier-free Guidance

The classifier-free guidance mechanism, as defined in Equation 5.6, is commonly used to guide the denoising process of a diffusion model. The model is commonly conditioned on a class label encoded as a label vector with 1024 values.

$$\tilde{\varepsilon}_\theta(z_\lambda, c) = \varepsilon_\theta(z_\lambda) + w\underbrace{(\varepsilon_\theta(z_\lambda, c) - \varepsilon_\theta(z_\lambda))}_{\text{class guidance}} \tag{5.6}$$

where $\tilde{\varepsilon}_\theta$ is the modified noise prediction, $\varepsilon_\theta(z_\lambda)$ the unconditional prediction, $\varepsilon_\theta(z_\lambda, c)$ the prediction conditioned on a label c and w a weight which controls the amount of guidance.

5.4.1.2 Multiple conditions Classifier-free guidance

As a stepping stone towards a privacy-preserving classifier-free guidance mechanism, we first explore modifying this mechanism for multiple conditions. This modification is formally defined in Equation 5.7.

$$\tilde{\varepsilon}_\theta(z_\lambda, c_1, c_2) = \varepsilon_\theta(z_\lambda) + w_1 \underbrace{(\varepsilon_\theta(z_\lambda, c_1) - \varepsilon_\theta(z_\lambda))}_{\text{class 1 guidance}} + w_2 \underbrace{(\varepsilon_\theta(z_\lambda, c_2) - \varepsilon_\theta(z_\lambda))}_{\text{class 2 guidance}} \tag{5.7}$$

Regularly, a diffusion model is trained both conditionally and unconditionally by randomly dropping the condition label during the training steps. We extend this behaviour so either condition can be removed randomly.

5.4.1.3 Classifier-free guidance with identity removal

Let I be a retrieval network that converts an image, x, to an identity embedding. Instead of conditioning the denoising process exclusively on a class label, we can instead condition on the combination of class label and an identity embedding which represents the identity of a real image, $I(x)$. This allows us to have an additional degree of freedom during sampling by controlling the guidance class-wise and identity-wise, as shown in Equation 5.8.

$$\tilde{\varepsilon}_\theta(z_\lambda, c, x) = \varepsilon_\theta(z_\lambda) + w_1 \underbrace{(\varepsilon_\theta(z_\lambda, c) - \varepsilon_\theta(z_\lambda))}_{\text{class guidance}} + w_2 \underbrace{(\varepsilon_\theta(z_\lambda, I(x)) - \varepsilon_\theta(z_\lambda))}_{\text{identity guidance}} \tag{5.8}$$

Taking into consideration the new sampling process, shown in Figure 5.5, there are multiple ways to potentially anonymize generated images:

1. Set $w_2 < 0$, and for each generated sample, choose an identity from the training set to steer away from. This is useful for the use case where we want to anonymize a single image.

2. Set $w_2 < 0$, and choose a random identity from the training set for each denoising step. While the previous situation is useful for a single image, by steering away from all identities in the training set, we can attempt to generate an image that is unlikely to belong to any existing patient.

3. Set $w_2 > 0$ and use a fixed identity embedding. Finally, this approach conditions all generated images on a single identity embedding, which may be obtained from a consenting patient. The largest downside of this approach is the possibility of losing some diversity in the synthetic dataset.

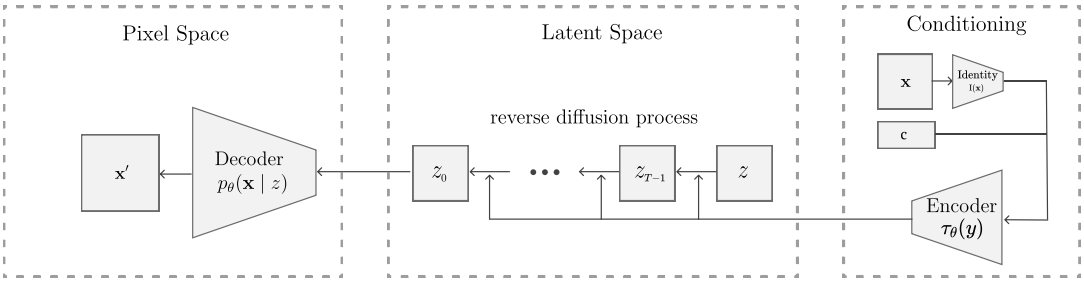


Figure 5.5: Overview of the sampling process for the classifier-free identity guidance mechanism. Starting from pure noise, z, a latent vector z_0 is generated conditioned on two conditions: the class c and an identity embedding.

5.4.2 Experiments

Conditioning on two conditions To evaluate this technique's viability, we perform experiments using Cardiomegaly as the first condition and Pleural Effusion as the second condition, as shown in Figure 5.6. We notice that it is possible to influence the generation of images by adjusting the secondary condition's weight up to some extent. One issue that we can immediately notice is the lack of granularity in changes: the generated images do not exhibit a smooth transformation as the value of w changes, instead suffering an abrupt change at an arbitrary w value. Despite this, it may still be possible to use this method for identity guidance, motivating the next experiment.

Conditioning on identity For this approach to work, we first need to decide on how to combine the time-step embedding, which encodes the current denoising time-step, the condition embedding, and the identity embedding. In the base case shown in Equation 5.6, we simply add both the time-step embedding and the condition embedding.

The condition embedding has size 1024, time-step embedding has size 1024 and the identity embedding has size 128. In our experiment, we concatenate both the condition embedding and identity embedding before adding both of them to the time-step embedding. In Figure 5.7, we can see an image showcasing how the use of an identity embedding, obtained from a test image on the left, impacts the generation of new images (on the right) given different w_2 weights. For the most part, the change in weight does not appear to have an important impact on the resemblance with the original image, leading us to believe that the generated images are merely arbitrary and are not being affected by the identity guidance system. We hypothesize that this experiment may have failed due to the amount of variance in the identity embeddings: while a class only takes two possible values for which we have a large number of examples, ideally, there are as many clusters of identity embeddings as there are patients and each of these patients only has a limited amount of available studies.

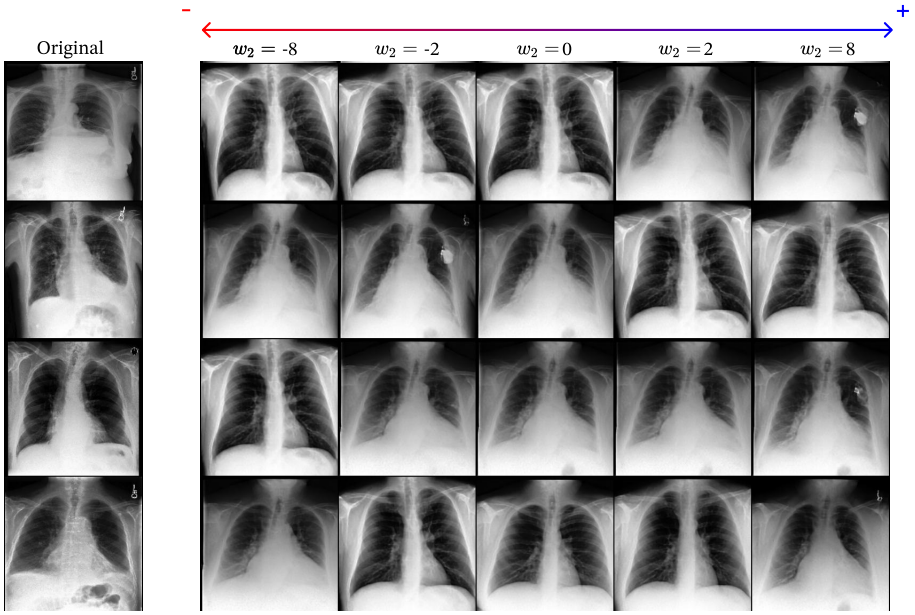


Figure 5.6: Generation of images using the classifier-free guidance mechanism for two conditions, Cardiomegaly and Pleural Effusion. The image on the left is the original image, and we use both condition labels to condition the generation. On the right, the images should exhibit the same Cardiomegaly label with varying strengths of Pleural Effusion, increasing from left to right.

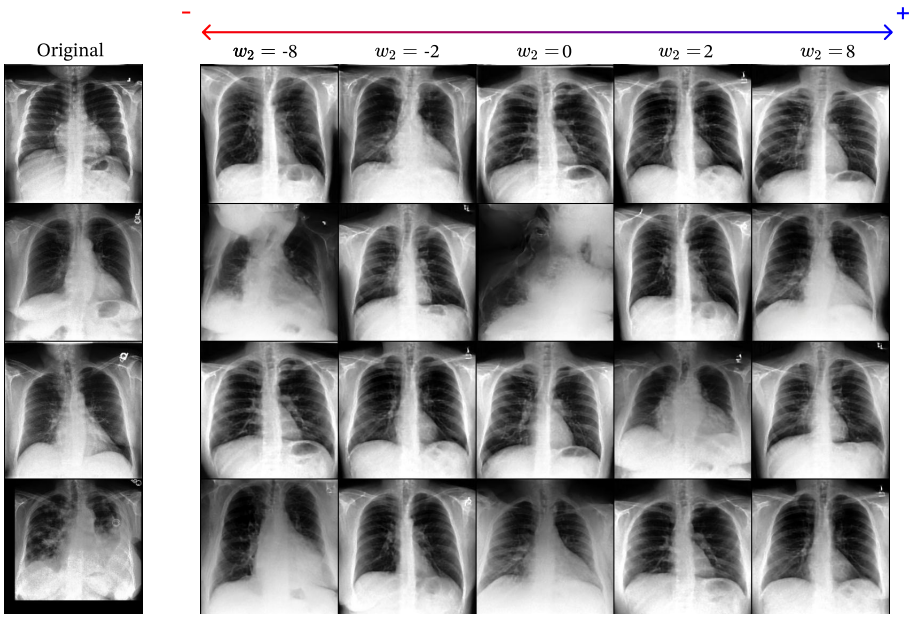


Figure 5.7: Generation of images using the classifier-free guidance mechanism for two conditions, Cardiomegaly and an identity embedding. The image on the left is the original image from which we use both condition's labels to condition the generation. On the right, the images should exhibit the same Cardiomegaly label with varying strengths of identity conditioning, increasing from left to right where the left should be the least similar and right most similar.

5.5 Summary

In this chapter, we have analyzed three approaches towards creating a privacy-preserving diffusion model. Yet, they achieved limited success when it comes to generating realistic and anonymous catalogues which could be used for case-based explanations. Both the DPDM and the Decoder-level anonymization can decrease the anonymization risk given an automated identity classifier. Their limited visual quality is still sub-par compared to real samples. In the following chapter, we analyze another anonymization technique that solves the aforementioned issue.

Chapter 6

Post-model Anonymization

6.1 Overview

The post-model anonymization procedure consists of multiple steps: image generation, patient retrieval and patient verification, which generate an anonymous synthetic dataset from which explanations can be retrieved, as shown in Figure 6.1.

This approach exhibits favourable characteristics, notably modularity, wherein each of its three distinct components can be substituted with an equivalent method. Additionally, it is firmly rooted in well-established methodologies documented in the existing literature. However, it is less efficient regarding computational power, as only a fraction of the synthetic images generated will be deemed anonymous. In contrast, the remaining non-anonymous images will effectively have been a waste of computing time.

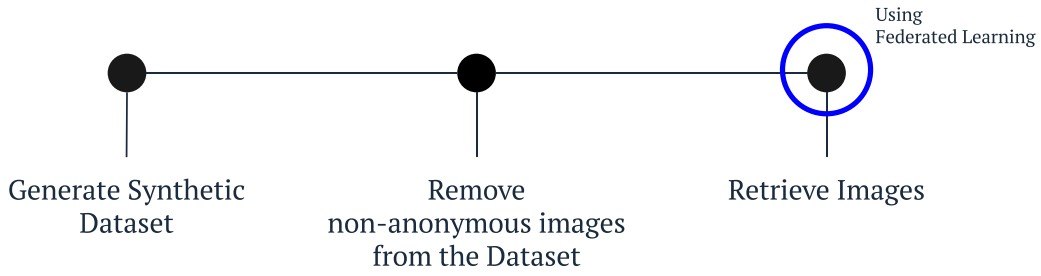


Figure 6.1: Post-model approach: first, generate synthetic images, then, anonymize them and retrieve them.

6.2 Method

Our methodology, summarized in Figure 6.2, follows the anonymization protocol defined by Packhäuser *et al.* [75]. Initially, we train a latent diffusion model and synthesize a dataset. Afterwards, a retrieval model for each image within this dataset will be used to identify which images in the training set closely resemble the synthetic image. Then, we employ an identity verification network to compare the anonymous image with its closest match and determine whether it likely

belongs to the same patient. This information allows us to remove non-anonymous images and, consequently, create an anonymous catalogue from which we can retrieve case-based explanations.

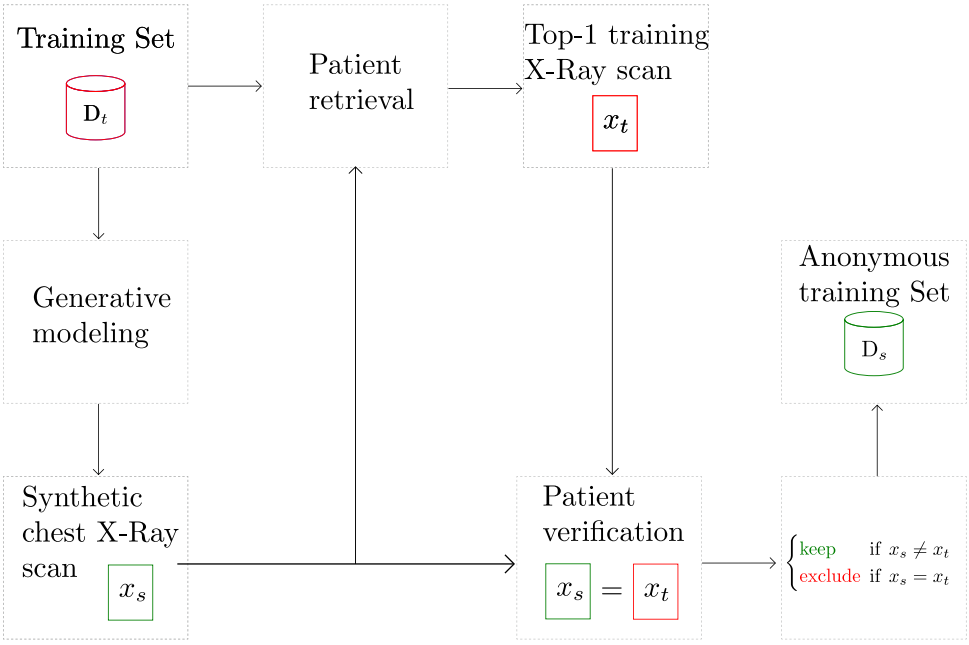


Figure 6.2: Anonymization Pipeline. A generative model generates synthetic samples based on the real training data. From these samples, we remove those closely resembling patients in the training set based on a patient retrieval and a patient verification model. This anonymous dataset serves as our knowledge base from which to retrieve explanations.

6.2.1 Image Generation using Latent Diffusion Models

We train a latent diffusion model which uses the Medfusion [68] architecture for image generation. Using it, we sample synthetic images to build a synthetic dataset that will be anonymized.

6.2.2 Retrieval and Verification Network

We need to identify if each synthetic image closely resembles any real patient in the training set. To do so, first, a patient retrieval network consists of a Siamese Neural Network [5] based on the ResNet-50 [30] architecture, which takes as input two images and, using a contrastive loss function, approximates embeddings so that the embeddings from the same patients are grouped together.

The verification model follows the same architecture but with a different task; instead of using a contrastive loss, this network compares both input images and classifies whether or not they belong to the same patient. For both models, the training data consists of positive training pairs obtained based on the patient identity information of a dataset and randomized negative pairs.

6.2.3 Anonymization Pipeline

Considering the synthetic dataset, the verification model, and the retrieval network, we have all the components required to obtain an anonymous dataset. The anonymization procedure we employ can be broken down into three separate steps:

1. **Compute training identity embeddings**: We calculate an identity embedding using the identity retrieval network for each training set image. These embeddings are stored in a KD-tree [3] to allow for a more efficient lookup step using a nearest-neighbour strategy later.

2. **Search nearest training image**: For each synthetic sample, we retrieve the top-1 most similar training sample. In this step, we compute the identity embedding of each image and perform a lookup of the previously mentioned KD-tree (Algorithm 1). Each real-synthetic image pair is stored in a list used in the next step.

Algorithm 1 Computing closest identity

Input: vector x, vector y_i, size m
Output: data closestVector
Initialize closestDistance $= \infty$.
Initialize closestVector $=$ None.
for $i = 0$ **to** $m-1$ **do**
  if $||x - y_i|| <$ closestDistance **then**
    closestDistance $= ||x - y_i||$
    closestVector $= y_i$
  end if
end for

3. **Remove non-anonymous image pairs**: For each real-synthetic image pair, we use the identity verification network to obtain a prediction of whether or not both images belong to the same patient. If this likelihood exceeds a predefined threshold, the synthetic image is deemed non-anonymous and removed from the synthetic dataset (Algorithm 2).

Algorithm 2 Filter dataset

Input: data x_i, data y_i, size m, model v, float t
for $i = 0$ **to** $m-1$ **do**
  if $v(x_i, y_i) > t$ **then**
    Delete x_i.
  end if
end for

Table 6.1: Quantitative evaluation of the images generated using the Medfusion model.

FID (↓)	P (↑)	R (↑)
62.25	68.00	22.17

6.3 Experiments

Data We perform experiments on the MIMIC-CXR-JPG dataset [41]. We chose to predict cardiomegaly as our classification task. From the available images, we select the ones from the posteroanterior (PA) view (96,161) and use the recommended data splits provided alongside the dataset.

Image Generation We generate MIMIC-CXR-JPG images using the Medfusion [68] architecture, the latent embeddings are encoded using a VAE, which accepts $3 \times 512 \times 512$ pixel images and has 8 embedding channels. The diffusion model uses DDIM [95], has label embedding and time embeddings of size 1024 and employs a scaled linear noise scheduler with $T = 1000$, which varies from $B_0 = 0.002, B_T = 0.02$. Using a U-Net [85] architecture, we train the model for a maximum of 1001 epochs with early stopping with patience of 30 epochs. We generate 1000 synthetic images for both datasets, with a balanced split between positive and negative cases.

Retrieval and Verification For the retrieval phase, similarly to Packhäuser *et al.* [75], we train our models for 30 epochs in the first phase, during which the model backbone is frozen, and 50 epochs during the second phase where all parameters are unfrozen. We use a learning rate of 0.158489 with weight decay of $1e^{-5}$, as suggested by the aper's authors and batch size 32. Our verification model was trained using the Adam optimizer [49] with $3 \times 256 \times 256$ images and a learning rate of 0.0001. We used a batch size of 32 and limited training to a maximum of 100 epochs with an early stopping criteria with 5 epoch patient. During the anonymization step, we considered images non-anonymous if the verification model prediction of a synthetic image and a real image belonging to the same patient was above a threshold of 0.5.

6.4 Results and discussion

Table 6.1 shows a quantitative image quality evaluation of the synthetic images generated by the Medfusion architecture. We report precision (P), recall (R), and the Fréchet inception distance (FID) [31]. To evaluate the image utility for classification tasks, we compare a classifier trained on synthetic data with a classifier trained on real data in Table 6.2. The classifier trained on synthetic data still achieves acceptable performance even though it has been trained on a comparatively small dataset. This indicates that the signal required to diagnose Cardiomegaly is still present in the new images.

Table 6.2: Classification performance reported for both the MIMIX-CXR-JPG and a synthetic dataset.

Dataset	F1 (↑)	ACC (↑)	Image Count
MIMIC-CXR-JPG	91.67	87.33	96,161
Synthetic	75.97	69.00	1,000

Both the verification and retrieval models showcase the ability to re-identify patients, as demonstrated by the metrics highlighted in Tables 6.3 and 6.4, respectively. For the verification model, we report the Area under the Receiver operating characteristic (AUC), precision (P), recall (R) and F1-score. For the retrieval task, we track the R-precision, a common information retrieval metric which is calculated as shown in Equation 6.1 using the number of relevant images retrieved (r) within the first R images, where R is the total number of relevant images existing in the dataset. Finally, the $mAP@R$, shown in Equation 6.2, is the mean of the average precision at R ($AP@R$) for each query.

$$\text{R-Precision} = \frac{r}{R} \tag{6.1}$$

$$mAP@R = \frac{1}{Q}\sum_{i=1}^{Q} AP_i@R \tag{6.2}$$

During the anonymization procedure, we removed all real-synthetic image pairs that likely belonged to a patient in the training set. This is determined by the verification model prediction, and some examples of image pairs at different prediction ranges are shown in Figure 6.3. As expected, the higher-value pairs look more similar than those within the lower ranges. During this process, we removed 161 images from the original synthetic dataset of 1,000 images leading to a final anonymized dataset featuring 839 images.

Finally, this method still presents some limitations. There is no upper bound on the number of synthetic images that will be removed, leading to a waste of computing power on generating images that will effectively be deleted. In future work, the anonymization mechanism could be integrated into the generative model to combat this issue. Another key issue is the empirical nature of this anonymization procedure, making it unproven. Unfortunately, the current best method to

Table 6.3: Quantitative evaluation of patient verification model.

AUC (↑)	ACC (↑)	F1 (↑)	P (↑)	R (↑)
93.72	95.32	90.92	90.77	91.07

Table 6.4: Quantitative evaluation of patient retrieval model.

mAP@R (↑)	P@1 (↑)	R_prec (↑)
79.32	94.27	80.87

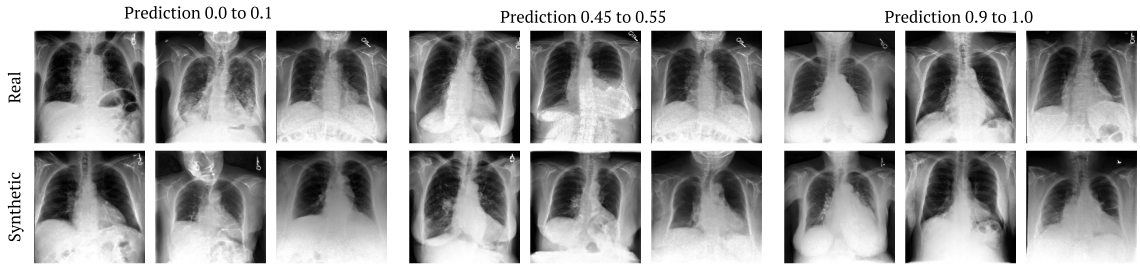


Figure 6.3: Different image pairs of anonymous images and their closest real counterpart in the training set, sorted by the predicted likelihood of belonging to the same patient. The images on the left are less likely to contain identifiable patient data, while images on the right are highly likely.

provide anonymization, with strong guarantees, is through Differential Privacy, which struggles to scale beyond low-resolution images, making it ineffective in generating explanations which are interpretable by humans.

6.5 Summary

We proposed a method to generate visually anonymous case-based explanations that retain their utility and realism by levering latent diffusion models. This solution demonstrates that the generated images are empirically unlikely to be traced back to the original patients, allowing for the visual explanation of classifier decisions without exposing patient data. This is a step towards integrating trustworthy and interpretable classifiers in the medical domain while preserving patient privacy.

Chapter 7

Case-based Explanation Retrieval

7.1 Overview

In previous chapters, we have analyzed different ways of generating anonymous images, which will serve as our explanation catalogues. This chapter presents methods for retrieving closely related anonymous images from said catalogues for a given test image. In Section 7.2, we describe how to do this in a centralized setting, while in Section 7.3, we extend this behaviour towards a de-centralized setting where multiple centres collaborate in both training the model and sharing synthetic explanations. Finally, in Section 7.4.2, we look over some safety considerations we must take when using the proposed solutions, and in Section 7.5, we draw some closing remarks about the chapter.

7.2 Retrieval in a Centralized setting

The simplest scenario for case-based explanation retrieval occurs when we consider a single centre. Here, for a test image, we need only classify it using a DL model and find, from the internal dataset, a set of training images which are closely related to the presented test image.

7.2.1 Method

As previously highlighted, the retrieval system should be based not on image similarity but on task-related features. For this purpose, we train a task classifier and use the features obtained by the said model as the basis for image comparison under the assumption that similar images will also have similar features.

7.2.2 Experiments

We employ a DenseNet121 [34] network trained using the Adam optimizer [49] with a learning rate of $1e^{-3}$ for up to 10 epochs with early stopping criteria and patience of 3 epochs. This model is tasked with a binary classification problem, determining whether a patient has or does not have

Table 7.1: Architecture of the DenseNet-121 [34] model. The features used for explanation retrieval obtained after the global average pool are highlighted in bold.

Layers	Output size	DenseNet-121
Convolution	$64 \times 128 \times 128$	7×7 conv, stride 2
Pooling	$64 \times 64 \times 64$	3×3 max pool, stride 2
Dense Block (1)	$256 \times 64 \times 64$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer (1)	$128 \times 64 \times 64$	1×1 conv
	$128 \times 32 \times 32$	2×2 average pool, stride 2
Dense Block (2)	$512 \times 32 \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Transition Layer (2)	$256 \times 32 \times 32$	1×1 conv
	$256 \times 16 \times 16$	2×2 average pool, stride 2
Dense Block (3)	$1024 \times 16 \times 16$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$
Transition Layer (3)	$512 \times 16 \times 16$	1×1 conv
	$512 \times 8 \times 8$	2×2 average pool, stride 2
Dense Block (4)	$1024 \times 8 \times 8$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 16$
Classification layer	$\mathbf{1024 \times 1 \times 1}$	7×7 global average pool Softmax layer

Cardiomegaly based on a chest X-ray image from MIMIC-CXR. To train it, we employ a Binary Cross Entropy loss function as defined in Equation 7.1.

$$\mathcal{L}_{\text{BCE}}(x,y) = \frac{1}{N}\sum_{n}^{N} -[y_n \cdot \log x_n + (1-y_n) \cdot \log(1-x_n)] \tag{7.1}$$

where x, y are the predicted and ground-truth labels, respectively, and N is the batch size.

To search for similar images, we use the last feature vector of the model, located before the dense classification layer and sized 1×1024 as shown in Table 7.1. To create our explanation catalogues, we used the set of 839 synthetic images generated in the previous chapter.

7.2.3 Results and discussion

Similar to previous studies [93], the explanation ranking is evaluated with the help of a radiologist. For this purpose, we perform two different experiments. For the first experiment, we used 5 test cases, each case containing a test image that we aim to diagnose and 10 synthetic catalogue images from which we will retrieve explanations. The second experiment differs by using a catalogue of 5 synthetic images and 5 real images to compare the utility of both image types. All the images were randomly sampled from the test set to ensure the classes were balanced. The catalogue size

is limited to 10 images due to the time-consuming nature of evaluating the images. Both the test sets [1] and the evaluation results [2] have been made openly available.

To evaluate ranking performance, we use a metric commonly used in ranking tasks, the normalized Discounted Cumulative Gain ($nDCG_p$) [21] metric (Equation 7.2) where p is the number of retrieved images. The relevance values for each image (rel_i) vary from 5.5, the most similar image, to 1.0, the least similar example, with a 0.5 step between each relevance value. And $IDCG_p$ is the ideal DCG_p value.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{7.2}$$

$$DCG_p = \sum_{i}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{7.3}$$

We also compare our CNN-based method against an approach which retrieves images with the highest Structural Similarity Index (SSIM), which compares the structure of two images as defined in Equation 7.4. This serves as a baseline to assess whether the proposed approach is necessary or not.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma xy + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{7.4}$$

The results of the ranking evaluation are shown in Figure 7.1. We can notice that our CNN-based method, on average, outperforms its SSIM counterpart, retrieving more relevant explanations. An example of the retrieval test we used to evaluate the solution is shown in Figure 7.2. In this example, the CNN-based retrieval mechanism obtained the same Top-3 images as the expert-based choices, although in a different order.

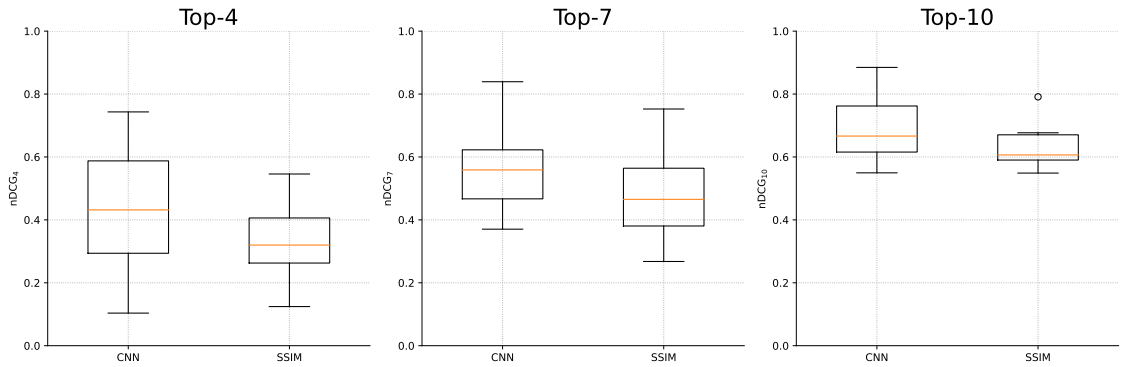


Figure 7.1: Boxplots regarding nDCG for two retrieval approaches, CNN and SSIM for the Top-4, Top-7 and Top-10 retrieved images.

[1] www.retrievalsets.dissertation.filipepcampos.com
[2] www.retrievalevaluation.dissertation.filipepcampos.com

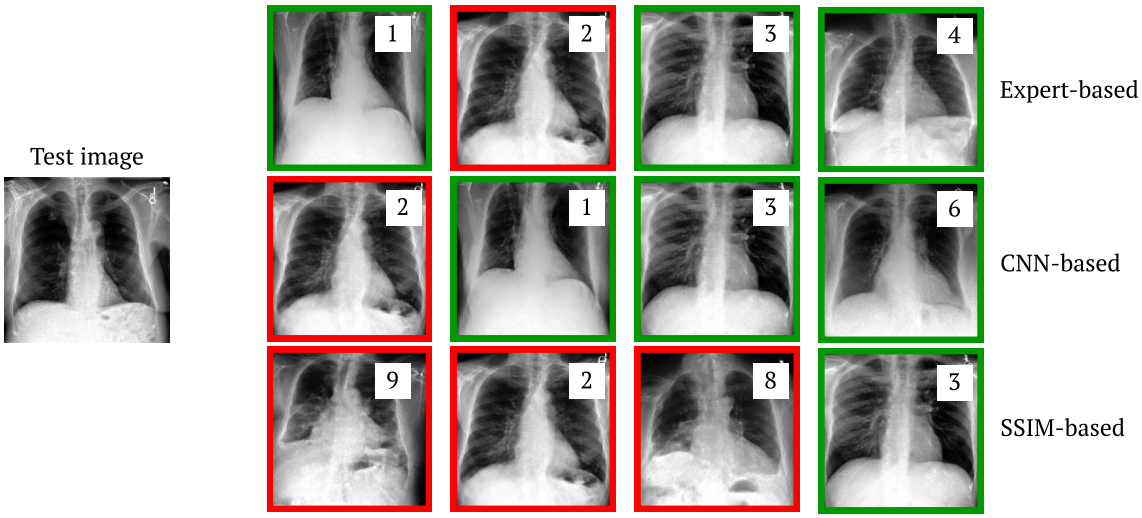


Figure 7.2: Example test image and corresponding Top-4 catalogue images retrieved by each corresponding method. The colour outline around each retrieved image indicates whether the test and catalogue images share the same class (green) or not (red). The ranking annotated on the top-right corner of each image indicates the ground truth based on expert rating.

Additionally, a radiologist assessed each image belonging to the test cases. In Table 7.2, we see that the generative model can conditionally generate images with a specific diagnosis while maintaining a similar class agreement compared to real images. We can also note that synthetic and real images are similarly relevant, supporting the hypothesis that the generated images can be used as alternatives to real ones.

7.3 Retrieval in a Federated-Learning setting

The main purpose of anonymization of medical case-based explanations is to enable their sharing for multiple purposes without disclosing personal patient information. One of the most useful applications for these explanations is sharing an anonymous catalogue across different medical institutions. Instead of using an explanation retrieval model trained in a centralized setting, we

Table 7.2: Label agreement between expert evaluation and the ground-truth label. A classification is correct if the image label and the radiologist agree on the diagnosis (Cardiomegaly or No Cardiomegaly). Cases where it is impossible to diagnose a sample confidently are deemed non-conclusive. We also report the average relevance of each type of image for the test cases with mixed image types.

Type	Correct	Incorrect	Non-conclusive	Average Relevance
Synthetic	70.67%	20.00%	9.33%	16.00 ± 4.30
Real	76.00%	20.00%	4.00%	16.50 ± 4.30

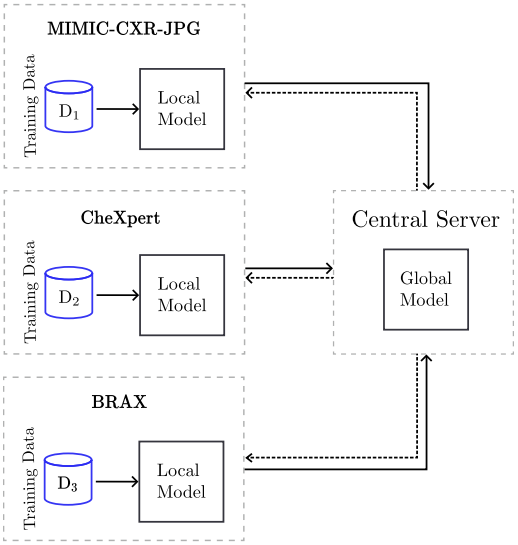


Figure 7.3: Overview of the Federated-learning architecture used. We simulate a Federated-Learning setting by introducing three distinct clients, each using a separate dataset. The training process is divided into rounds where, in each round, a client trains a single epoch locally and submits the updated local weights to the central server. At the end of each round the server aggregates all the weights using the FedAvg strategy and relays back the updated weights to the local clients.

can use a model in a federated learning setting where multiple clients (hospitals or institutions) collaboratively contribute towards the training. Figure 7.3 shows an overview of this method.

7.3.1 Method

To perform federated learning, we require multiple clients to train a model on their own data and a server that aggregates the results obtained by the clients and relays the new model parameters back. For this work, we used the FedAvg [61], presented in Chapter 2 as the aggregation technique. More specifically, we use the implementation provided by the Flower [4] package.

7.3.2 Experiments

To simulate an FL scenario we utilize three datasets, one for each of 3 centres, MIMIC-CXR-JPG [41], CheXpert [35] and BRAX [83]. To obtain a baseline performance on how our model would perform in a centralized setting, we first re-trained the architecture proposed in Section 7.2 and tested each model across the three datasets.

Similarly to the centralized setting, we employ a DenseNet121 [34] network trained using the Adam optimizer [49] with a learning rate of $1e^{-4}$. We perform 10 rounds of FL using the FedAvg strategy, sample from all clients for both fit and evaluation and test the results on each individual dataset.

7.3.3 Results and discussion

The results of the centralized performance can be seen in Table 7.3; as we are dealing with a regular classification problem, we report the Area under the ROC Curve (AUC), F1-Score (F1) and accuracy (ACC). In particular, we focus on the AUC, using it to evaluate the model's performance. In Table 7.4, compare our FL solution against the centrally trained models and notice that this proposed model outperforms the centralized solutions for both the MIMIC and CheXpert dataset while being very competitive with the BRAX model.

Table 7.3: Performance of the DenseNet121 model on a centralized setting.

Dataset	AUC (↑)	F1 (↑)	ACC (↑)
MIMIC-CXR-JPG	88.60	89.63	84.62
CheXpert	68.52	27.50	75.21
BRAX	92.00	49.32	91.72

Table 7.4: Comparison in performance, measured using AUC, of models trained on different datasets and a Federated-Learning approach.

	Test Dataset		
Train Dataset	MIMIC-CXR-JPG	CheXpert	BRAX
MIMIC-CXR-JPG	88.60	71.63	86.34
CheXpert	79.03	68.52	75.18
BRAX	87.40	64.68	**92.00**
All (FL)	**89.48**	**78.64**	91.09

Recalling the evaluation procedure used in Section 7.3.3, we want to ensure that the retrieval model trained under FL is capable of matching the performance of the previously studied centralized approach. For this, for each of the 10 test cases we simply ranked all the catalogue images and compared the performance against the existing models. In Figure 7.4 we can see that the new model has a comparable performance to the centralized CNN for the presented test scenarios while achieving a unified feature representation for multiple datasets.

Figure 7.5 showcases some test examples for which we retrieved case-based explanations using the proposed FL model. These explanations were retrieved from the combination of the Synthetic MIMIC-CXR-JPG, CheXpert and BRAX datasets.

7.4 Practical Considerations

In implementing retrieval methods for anonymized case-based explanations, some practical considerations must be addressed to ensure the methods' effectiveness and safety in real-world medical settings. This section discusses these considerations, namely a possible method for model deployment and an overview of safety considerations, taking different usage scenarios into consideration.

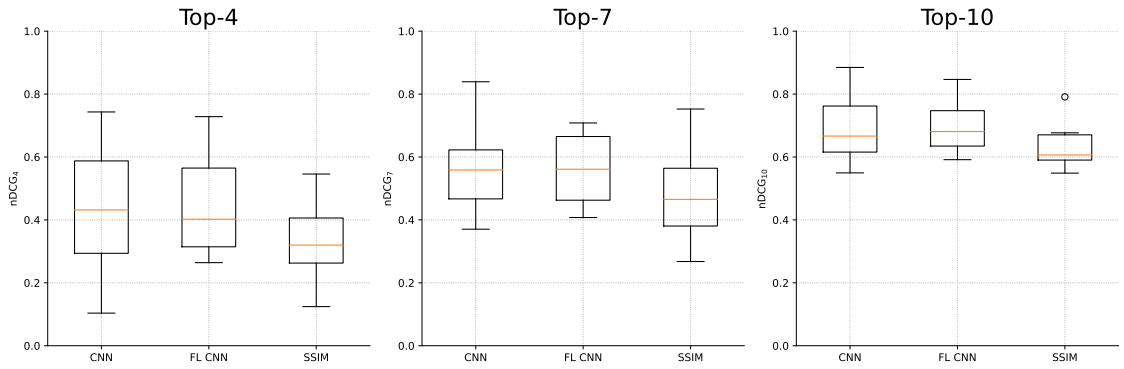


Figure 7.4: Boxplots regarding nDCG for three retrieval approaches, CNN, CNN trained using FL and SSIM for the Top-4, Top-7 and Top-10 retrieved images.

7.4.1 Model Usage

At this point, we have a usable FL retrieval model to obtain feature-related embeddings for any centre or client input images. An important step towards its use in a real scenario would be defining how it can be used in practice. Even though the catalogue images have been anonymized, it is still ideal to minimize unnecessary data sharing, allowing each centre closer control of its data. For this, we propose creating a centralized database of embeddings, which any client can later query. To do so, each centre should compute an image embedding for each of their anonymized images and submit it, alongside a randomized identifier, to the centralized server as demonstrated in Figure 7.6.

After the database has been created, each client can query the server by submitting a retrieval embedding to obtain the location of the closest explanations and request them. This procedure is shown in Figure 7.7. This mechanism allows each centre to only share the explanations immediately required by another centre without needing to trust the central server.

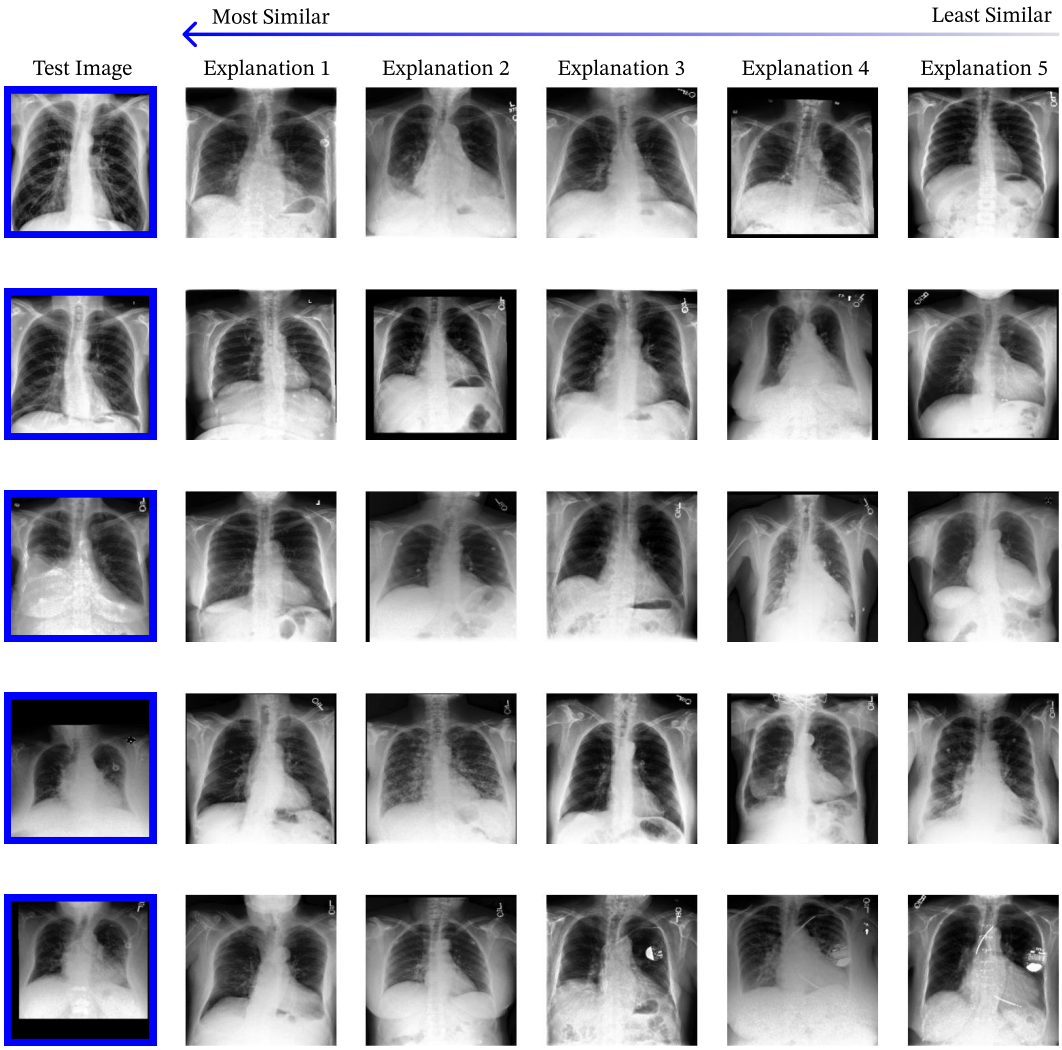


Figure 7.5: Example of images retrieved using our FL retrieval model as explanations for multiple test images. The images are ordered from most to least similar and were retrieved from the combination of three datasets: MIMIC-CXR-JPG, CheXpert and BRAX.

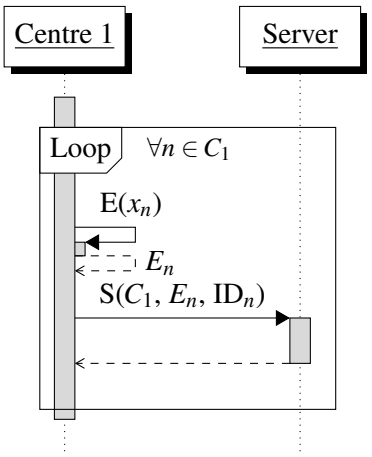


Figure 7.6: Aggregation mechanism to build a centralized embedding database. Each client computes an embedding E_n for each anonymous image x_n using the retrieval model E. Each of these embeddings is then submitted to the server alongside a randomly generated identifier, ID$_n$.

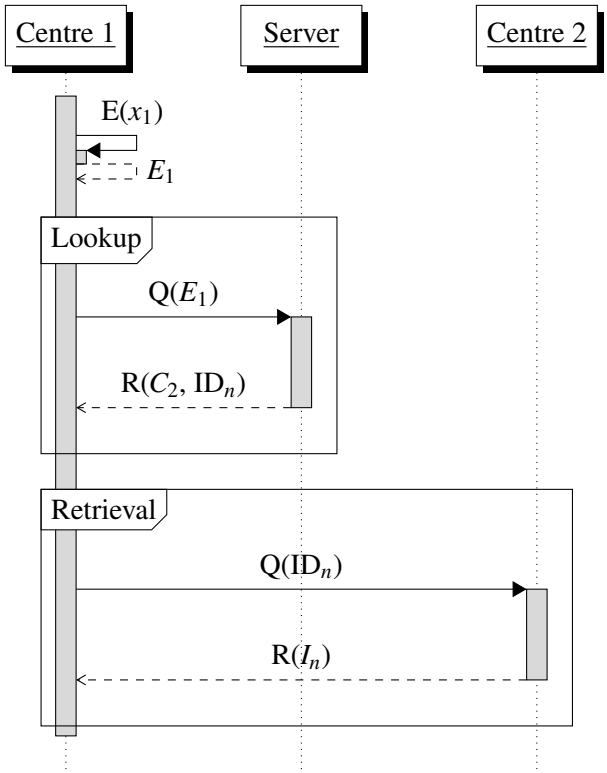


Figure 7.7: Example of a retrieval mechanism under a federated learning setting. Centre 1 uses the retrieval model to obtain an embedding, E_1 representing the test image x_1. Afterwards, it queries the server, which looks up its vector database, to find the location of the closest explanation, owned by C_2 and identified by ID$_n$. Finally, the client retrieves the explanation from the second centre. This mechanism can easily be expanded to retrieve multiple explanations per test image.

7.4.2 Safety

Returning to the topic of safety, discussed in Chapter 3.5, this solution can be analyzed under different scenarios. Each scenario can have different types of attack vectors such as **access to samples of the case-based explanations**, **black-box access** to a model where an adversary only has access to an API or **white-box access** where someone can analyze the model weights. Table 7.5 shows an overview of each setting's attack vectors.

Table 7.5: Overview of possible attack methods for different usage settings for the Case-based explanation retrieval system.

Method	Explanation Access	Black-Box Setting	White-Box Setting
Centralized (Internal use)	✓	-	-
Centralized (External use)	✓	✓	-
Federated-Learning	✓	✓	✓

7.4.2.1 Centralized

In a centralized setting, the only component that can be shared is the explanation catalogue. Considering this, the risk of re-identification can be estimated based on the performance of the chosen anonymization method. For instance, we can employ the post-model anonymization technique analyzed during Chapter 6, whose re-identification risk is determined by the performance of both the identity retrieval and identity verification models. Suppose the model is an openly available API whose input is a test image and output is a set of explanations. In that case, the risk of performing a black-box attack such as membership inference to determine whether a certain sample belongs to its training data is limited since potential attackers would not have direct access to model predictions.

7.4.2.2 Federated-Learning

A Federated-Learning model will present the same attack vectors as the centralized model and the risk of sharing model weights amongst centres. This weight sharing makes the retrieval model susceptible to membership inference white-box attacks unless all the involved centres are trustworthy. This risk can be negated by introducing instance-level differential privacy techniques [67]. Since this issue has been previously explored [67] it should be easily integrated into the existing workflow. There is the additional risk of a centre purposefully interfering with the training of the shared model, yet this would not have an important impact on patient privacy; instead, the model's training capabilities would be limited until said centre is removed from the cluster. In practice, the system could still be used for inference, assuming each centre uses a frozen version of the retrieval model and only updates its weights at large yet regular intervals.

7.5 Summary

This chapter focused on the methods for retrieving anonymized case-based explanations for medical images. In a centralized setting, the retrieval process involved using a deep learning (DL) model to classify a test image and then identifying closely related training images from an internal dataset. This method effectively generated relevant and accurate case-based explanations based on task-related features extracted by the model. The DenseNet121 architecture, used for feature extraction, facilitated the retrieval of highly relevant images to the given test image, as demonstrated by the improved nDCG scores compared to a baseline technique, which employed a SSIM metric.

Extending the retrieval method to a federated learning (FL) setting allowed multiple institutions to collaborate in training the model while maintaining data privacy. Each institution contributes to a shared model without directly exposing its local data. The federated approach maintained the performance benefits observed in the centralized setting, as indicated by comparable nDCG scores. Additionally, the federated model enabled the retrieval of explanations from a more diverse group of datasets, enhancing the robustness and generalizability of the retrieved explanations. Finally, we addressed some practical considerations for deploying the retrieval models, which are crucial for ensuring that the retrieval methods can be safely implemented in real-world medical environments.

Using centralised and federated learning approaches, we established effective retrieval methods for anonymized case-based explanations. These methods provide a robust framework for generating and sharing medical image explanations while adhering to stringent privacy requirements, thus facilitating the broader adoption of deep learning techniques in the medical field.

Chapter 8

Conclusion

Over the years, deep learning classifiers have been applied with great success in medical diagnosis tasks, with the aim of improving the efficiency of medical workflows. Yet, due to the limited understanding of the rationale behind automated diagnosis, these systems are not trustworthy enough to truly integrate the medical professional's toolsets. This dissertation explored a method of solving this issue by enhancing the interpretability of DL models through the use of anonymous case-based explanations that explain decisions using visual examples.

Given the sensitive nature of medical images, the images used for case-based explanations must first be anonymized to ensure compliance with regulations such as the GDPR and HIPAA and upkeeping the patient's essential right to privacy. After this anonymization procedure, these images can be shared between professionals or even amongst different hospitals in a Federated Learning setting.

To perform anonymization, this work focused on the use of diffusion models, which, in recent years, have excelled in image generation tasks, a task strongly related to visual anonymization. We also explored how these models can be applied to medical use cases to generate synthetic medical images.

This dissertation unifies multiple research areas, such as diffusion models, anonymization, and explanation retrieval, into a single system which can explain a classifier's decisions through the use of anonymous case-based explanations. For this purpose, we propose two different approaches, which have been thoroughly explored during the course of this work: post-model anonymization and in-model anonymization. Additionally, we evaluate the effectiveness of the proposed methods with the aid of an experienced radiologist.

This work allowed us to answer the research questions proposed in Chapter 1.3:

RQ.1 Can anonymous case-based explanations be generated using diffusion models? Yes. During this work, we have explored multiple methods capable of generating anonymous medical images. These images have undergone thorough evaluation by an experienced radiologist to ensure their effectiveness and reliability.

RQ.2 Is it possible to retrieve the aforementioned explanations in a federated learning setting? Yes. We have successfully created a retrieval model that operates under a Federated Learning setting, demonstrating its feasibility and effectiveness in this distributed framework. This model performs competitively against a single-centre solution while achieving a unified feature representation which can be used to retrieve images from multiple centres.

RQ.3 Can the de-anonymization risk associated with the aforementioned anonymous explanations be evaluated and quantified? Yes. In the anonymization proposed, post-model anonymization, the risk of de-anonymization is determined based on the performance of both the patient verification network and the patient retrieval network.

Even though this dissertation has met its objectives, moving forward there are many avenues to further explore this topic and build upon the existing work:

- **Differentially Private Federated-Learning Retrieval Model** - Currently, the retrieval model trained under FL is susceptible to malicious clients which have access to the weights of the shared model, which has been trained using real data from multiple centres. Future work should focus on integrating differential privacy techniques to mitigate this risk.
- **Different Modalities** - The proposed solution is task-agnostic. Therefore, it should be directly applicable to different medical modalities, such as CT scans and MRI. Future work should take the opportunity to evaluate how well the proposed solution works across multiple of these modalities, therefore evaluating how adaptable the methods are.
- **Improved retrieval methods** - As shown in Chapter 3.4, multiple works improve upon the retrieval method we suggested as a baseline. Future research could focus on evaluating the performance of different methods, such as Interpretability-Guided Content-Based Medical Image Retrieval (IG-CBIR) [93], in retrieving synthetic medical images.
- **Counterfactual explanations** - Exploring counterfactual explanations could provide additional insights and improve the interpretability of DL models in medical diagnosis tasks. Future work should investigate the integration of counterfactual explanations into the proposed system to further enhance its explanatory power.

This dissertation has laid a robust foundation for generating and utilizing anonymous case-based explanations in medical diagnostics. By effectively addressing privacy concerns and maintaining high interpretability, the proposed methods facilitate the broader adoption of deep learning techniques in the medical field. Moreover, the exploration of federated learning frameworks opens up new possibilities for collaborative, privacy-preserving medical research and diagnostics.

In conclusion, this work not only advances the field of medical image analysis but also paves the way towards practical solutions for real-world implementation. By ensuring both privacy and interpretability, these contributions lead to a more trustworthy and widespread use of deep learning in healthcare.

References

[1] Amol Khanna, Vincent Schaffer, Gamze Gursoy, and Mark Gerstein. Privacy-preserving Model Training for Disease Prediction Using Federated Learning with Differential Privacy. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, July 2022. Cited on page 12.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. arXiv:1701.07875. Cited on page 26.

[3] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, sep 1975. Cited on page 47.

[4] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Hei Li Kwing, Titouan Parcollet, Pedro PB de Gusmão, and Nicholas D Lane. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020. Cited on page 55.

[5] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, page 737–744, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc. Cited on page 46.

[6] Filipe Campos, Liliana Petrychenko, Luís F. Teixeira, and Wilson Silva. Latent diffusion models for privacy-preserving medical case-based explanations. In *submitted to EXPLIMED, ECAI 2024*. Cited on page 3.

[7] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting Training Data from Diffusion Models, January 2023. arXiv:2301.13188. Cited on pages 29 and 30.

[8] Jiawei Chen, Janusz Konrad, and Prakash Ishwar. VGAN-Based Image Representation Learning for Privacy-Preserving Facial Expression Recognition, September 2018. arXiv:1803.07100. Cited on page 25.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, June 2020. arXiv:2002.05709. Cited on page 11.

[10] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning, November 2020. arXiv:2011.10566. Cited on page 11.

[11] Zhiguang Chu, Jingsha He, Dongdong Peng, Xing Zhang, and Nafei Zhu. Differentially Private Denoise Diffusion Probability Models. *IEEE Access*, 11:108033–108040, 2023. Cited on page 22.

[12] Council of European Union. Council regulation (EU) no 679/2016. Online at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504, 2016. Cited on pages 4 and 27.

[13] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, September 2023. Cited on page 7.

[14] Pranjit Das and Arambam Neelima. An overview of approaches for content-based medical image retrieval. *International Journal of Multimedia Information Retrieval*, 6(4):271–280, December 2017. Cited on page 24.

[15] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis, June 2021. arXiv:2105.05233. Cited on page 15.

[16] Zhitong Ding, Shuqi Jiang, and Jingya Zhao. Take a close look at mode collapse and vanishing gradient in GAN. In *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, pages 597–602, May 2022. Cited on page 7.

[17] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially Private Diffusion Models, August 2023. arXiv:2210.09929. Cited on pages 22 and 35.

[18] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning, March 2017. arXiv:1702.08608. Cited on page 23.

[19] Qi Dou, Tiffany Y. So, Meirui Jiang, Quande Liu, Varut Vardhanabhuti, Georgios Kaissis, Zeju Li, Weixin Si, Heather H. C. Lee, Kevin Yu, Zuxin Feng, Li Dong, Egon Burian, Friederike Jungmann, Rickmer Braren, Marcus Makowski, Bernhard Kainz, Daniel Rueckert, Ben Glocker, Simon C. H. Yu, and Pheng Ann Heng. Federated deep learning for detecting covid-19 lung abnormalities in ct: a privacy-preserving multinational validation study. *npj Digital Medicine*, 4(1):60, Mar 2021. Cited on page 11.

[20] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks?, 2023. arXiv:2302.01316. Cited on pages 29 and 30.

[21] Kelwin Fernandes and Jaime S. Cardoso. Hypothesis transfer learning based on structural model similarity. *Neural Computing and Applications*, 31(8):3417–3430, Aug 2019. Cited on page 53.

[22] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. Large-scale privacy protection in Google Street View. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2373–2380, September 2009. Cited on page 21.

[23] Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L. Smith, Olivia Wiles, and Borja Balle. Differentially Private Diffusion Models Generate Useful Synthetic Images, February 2023. arXiv:2302.13861. Cited on page 22.

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, June 2014. arXiv:1406.2661. Cited on page 7.

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised Learning, September 2020. arXiv:2006.07733. Cited on page 11.

[26] Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney. Integrating Utility into Face De-identification. In George Danezis and David Martin, editors, *Privacy Enhancing Technologies*, Lecture Notes in Computer Science, pages 227–242, Berlin, Heidelberg, 2006. Springer. Cited on page 21.

[27] Ralph Gross, Latanya Sweeney, Jeffrey Cohn, Fernando de la Torre, and Simon Baker. Face De-identification. In Andrew Senior, editor, *Protecting Privacy in Video Surveillance*, pages 129–146. Springer, London, 2009. Cited on page 21.

[28] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5):1–42, September 2019. Cited on page 23.

[29] Corentin Hardy, Erwan Le Merrer, and Bruno Sericola. MD-GAN: Multi-Discriminator Generative Adversarial Networks for Distributed Datasets. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 866–877, May 2019. arXiv:1811.03850. Cited on page 11.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. Cited on pages 22 and 46.

[31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. arXiv:1706.08500. Cited on page 48.

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. Cited on pages 7, 8, 9, 10, and 15.

[33] Hailong Hu and Jun Pang. Membership Inference of Diffusion Models, January 2023. arXiv:2301.09956. Cited on pages 29 and 30.

[34] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269. IEEE, 2017. Cited on pages 51, 52, and 55.

[35] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A

Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, January 2019. arXiv:1901.07031. Cited on pages 3, 32, 33, and 55.

[36] Abdul Jabbar, Xi Li, and Bourahla Omar. A Survey on Generative Adversarial Networks: Variants, Applications, and Training, June 2020. arXiv:2006.05132. Cited on page 7.

[37] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion Models for Counterfactual Explanations, March 2022. arXiv:2203.15636. Cited on page 16.

[38] Yeon Uk Jeong, Soyoung Yoo, Young-Hak Kim, and Woo Hyun Shim. De-Identification of Facial Features in Magnetic Resonance Images: Software Development Using Deep Learning Technology. *Journal of Medical Internet Research*, 22(12):e22739, December 2020. Cited on page 5.

[39] Zhanglong Ji, Zachary C. Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review, 2014. arXiv:1412.7584. Cited on page 35.

[40] Alistair E. W. Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. The mimic-cxr database, 2019. URL: https://physionet.org/content/mimic-cxr/. Cited on page 33.

[41] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, November 2019. arXiv:1901.07042. Cited on pages 3, 32, 33, 48, and 55.

[42] Fiona Victoria Stanley Jothiraj and Afra Mashhadi. Phoenix: A federated generative diffusion model, 2023. arXiv:2306.04098. Cited on page 11.

[43] Georgios Kaissis, Marcus R. Makowski, Marcus R. Makowski, Daniel Rueckert, Daniel Rückert, and Rickmer Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, June 2020. Cited on page 12.

[44] Georgios Kaissis, Alexander Ziller, Jonathan Passerat–Palmbach, Jonathan Passerat-Palmbach, Theo Ryffel, Théo Ryffel, Théo Ryffel, Dmitrii Usynin, Dmitrii Usynin, Andrew Trask, Ionésio Da Lima, Ionésio Lima, Jason Mancuso, Jason Mancuso, Friederike Jungmann, Marc Steinborn, Marc-Matthias Steinborn, Andreas Saleh, Andreas Saleh, Marcus R. Makowski, Marcus R. Makowski, Daniel Rueckert, and Rickmer Braren. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, 3(6):473–484, May 2021. Cited on page 11.

[45] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks, March 2019. arXiv:1812.04948. Cited on page 18.

[46] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion Models for Medical Image Analysis: A Comprehensive Survey, June 2023. arXiv:2211.07804. Cited on pages 7 and 17.

[47] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan. Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review. *Journal of Imaging*, 9(4), 2023. Cited on page 7.

[48] Matthias Keicher, Matan Atad, David Schinz, Alexandra S. Gersing, Sarah C. Foreman, Sophia S. Goller, Juergen Weissinger, Jon Rischewski, Anna-Sophia Dietrich, Benedikt Wiestler, Jan S. Kirschke, and Nassir Navab. Semantic Latent Space Regression of Diffusion Autoencoders for Vertebral Fracture Grading, March 2023. arXiv:2303.12031. Cited on page 19.

[49] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. arXiv:1412.6980. Cited on pages 39, 48, 51, and 55.

[50] Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. arXiv:1906.02691. Cited on page 7.

[51] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, December 2022. arXiv:1312.6114. Cited on page 7.

[52] Marvin Klemp, Kevin Rösch, Royden Wagner, Jannik Quehl, and Martin Lauer. LDFA: Latent Diffusion Face Anonymization for Self-Driving Applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3198–3204, 2023. Cited on page 22.

[53] Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An Efficient Membership Inference Attack for the Diffusion Model by Proximal Initialization, October 2023. arXiv:2305.18355. Cited on pages 29 and 30.

[54] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion Models already have a Semantic Latent Space, March 2023. arXiv:2210.10960. Cited on pages 18 and 20.

[55] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks, April 2020. arXiv:1812.06127. Cited on page 12.

[56] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair Resource Allocation in Federated Learning, February 2020. arXiv:1905.10497. Cited on page 12.

[57] Wei Li, Jinlin Chen, Zhenyu Wang, Zhidong Shen, Chao Ma, and Xiaohui Cui. IFL-GAN: Improved Federated Learning Generative Adversarial Network With Maximum Mean Discrepancy Model Aggregation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10502–10515, December 2023. Cited on page 11.

[58] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion Models for Image Restoration and Enhancement – A Comprehensive Survey, August 2023. arXiv:2308.09388. Cited on page 7.

[59] Lorenzo Luzi, Ali Siahkoohi, Paul M. Mayer, Josue Casco-Rodriguez, and Richard Baraniuk. Boomerang: Local sampling on image manifolds using diffusion models, October 2022. arXiv:2210.12100. Cited on page 22.

[60] Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership Inference Attacks against Diffusion Models, March 2023. arXiv:2302.03262. Cited on pages 29 and 30.

[61] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data, October 2016. arXiv:1602.05629. Cited on pages 11, 12, and 55.

[62] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019. Cited on page 23.

[63] Grégoire Montavon, Sebastian Bach, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65:211–222, May 2017. arXiv:1512.02479. Cited on pages 25 and 27.

[64] H. Montenegro, W. Silva, and J. S. Cardoso. Towards Privacy-preserving Explanations in Medical Image Analysis, July 2021. arXiv:2107.09652. Cited on page 25.

[65] Helena Montenegro, Wilson Silva, and Jaime S. Cardoso. Privacy-Preserving Generative Adversarial Network for Case-Based Explainability in Medical Image Analysis. *IEEE Access*, 9:148037–148047, 2021. Cited on page 26.

[66] Helena Montenegro, Wilson Silva, Alex Gaudio, Matt Fredrikson, Asim Smailagic, and Jaime S. Cardoso. Privacy-Preserving Case-Based Explanations: Enabling Visual Interpretability by Protecting Privacy. *IEEE Access*, 10:28333–28347, 2022. Cited on pages 30 and 38.

[67] Laura Latorre Moreno. Towards case-based interpretability for federated learning models, 2023. Cited on pages 11, 25, and 60.

[68] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, Jakob Nikolas Kather, and Daniel Truhn. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, July 2023. Cited on pages 18, 38, 46, and 48.

[69] E.M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, February 2005. Cited on pages 21 and 26.

[70] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, Jul 2022. Cited on page 32.

[71] Alex Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models, February 2021. arXiv:2102.09672. Cited on pages 8, 10, and 15.

[72] J. R. Norris. *Discrete-time Markov chains*, page 1–59. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997. doi:10.1017/CBO9780511810633.003. Cited on page 8.

[73] OMG. Business Process Model and Notation (BPMN), Version 2.0, January 2011. URL: http://www.omg.org/spec/BPMN/2.0. Cited on page 28.

[74] Muhammad Owais, Muhammad Arsalan, Jiho Choi, and Kang Ryoung Park. Effective Diagnosis and Treatment through Content-Based Medical Image Retrieval (CBMIR) by Using Artificial Intelligence. *Journal of Clinical Medicine*, 8(4):462, April 2019. Cited on page 24.

[75] Kai Packhäuser, Lukas Folle, Florian Thamm, and Andreas Maier. Generation of Anonymous Chest Radiographs Using Latent Diffusion Models for Training Thoracic Abnormality Classification Systems, November 2022. arXiv:2211.01323. Cited on pages 22, 23, 39, 45, and 48.

[76] Kai Packhäuser, Sebastian Gündel, Nicolas Münster, Christopher Syben, Vincent Christlein, and Andreas Maier. Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data. *Scientific Reports*, 12(1):14851, September 2022. Cited on page 30.

[77] Luis Perez and Jason Wang. The Effectiveness of Data Augmentation in Image Classification using Deep Learning, December 2017. arXiv:1712.04621. Cited on page 11.

[78] Walter H. L. Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F. da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Brain Imaging Generation with Latent Diffusion Models, September 2022. arXiv:2209.07162. Cited on page 17.

[79] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10609–10619, New Orleans, LA, USA, June 2022. IEEE. Cited on pages 18 and 19.

[80] Elizabeth Puddy and Catherine Hill. Interpretation of the chest radiograph. *Continuing Education in Anaesthesia Critical Care & Pain*, 7(3):71–75, 05 2007. Cited on page 33.

[81] Gérald Quatrehomme, Stéphane Cotin, Gérard Subsol, Hervé Delingette, Yves Garidel, Gilles Grévin, Márta Fidrich, Paul Bailet, and Amédée Ollier. A Fully Three-Dimensional Method for Facial Reconstruction Based on Deformable Models. *Journal of Forensic Sciences*, 42(4):649, 1997. Cited on page 5.

[82] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive Federated Optimization, September 2021. arXiv:2003.00295. Cited on page 12.

[83] Eduardo P. Reis, Joselisa P. Q. de Paiva, Maria C. B. da Silva, Guilherme A. S. Ribeiro, Victor F. Paiva, Lucas Bulgarelli, Henrique M. H. Lee, Paulo V. Santos, Vanessa M. Brito, Lucas T. W. Amaral, Gabriel L. Beraldo, Jorge N. Haidar Filho, Gustavo B. S. Teles, Gilberto Szarf, Tom Pollard, Alistair E. W. Johnson, Leo A. Celi, and Edson Amaro. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487, Aug 2022. Cited on pages 3, 32, 33, and 55.

[84] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. Cited on pages 8 and 16.

[85] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. arXiv:1505.04597. Cited on pages 10, 19, 36, and 48.

[86] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement, June 2021. arXiv:2104.07636. Cited on page 17.

[87] C.G. Schwarz, W.K. Kremers, A. Arani, M. Savvides, R.I. Reid, J.L. Gunter, M.L. Senjem, P.M. Cogswell, P. Vemuri, K. Kantarci, D.S. Knopman, R.C. Petersen, C.R. Jack, Jr., and Alzheimer's Disease Neuroimaging Initiative the. A face-off of MRI research sequences by their need for de-facing. *NeuroImage*, 276, 2023. Cited on page 5.

[88] Christopher G. Schwarz, Walter K. Kremers, Val J. Lowe, Marios Savvides, Jeffrey L. Gunter, Matthew L. Senjem, Prashanthi Vemuri, Kejal Kantarci, David S. Knopman, Ronald C. Petersen, and Clifford R. Jack. Face recognition from research brain PET: An unexpected PET problem. *NeuroImage*, 258:119357, September 2022. Cited on page 5.

[89] Kenneth P. Seastedt, Patrick Schwab, Zach O'Brien, Edith Wakida, Karen Herrera, Portia Grace F. Marcelo, Louis Agha-Mir-Salim, Xavier Borrat Frigola, Emily Boardman Ndulue, Alvin Marcelo, and Leo Anthony Celi. Global healthcare fairness: We should be sharing more, not less, data. *PLOS Digital Health*, 1(10):e0000102, October 2022. Cited on page 5.

[90] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-Institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. *Brainlesion*, 11383:92–104, January 2019. Cited on page 11.

[91] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE Computer Society, May 2017. Cited on pages 28 and 29.

[92] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, July 2019. Cited on page 11.

[93] Wilson Silva, Alexander Poellinger, Jaime S. Cardoso, and Mauricio Reyes. Interpretability-guided content-based medical image retrieval. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 305–314, Cham, 2020. Springer International Publishing. Cited on pages 24, 52, and 63.

[94] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. Cited on page 7.

[95] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models, October 2022. arXiv:2010.02502. Cited on pages 8, 10, and 48.

[96] Sunghwan Moon and Won Hee Lee. Privacy-Preserving Federated Learning in Healthcare. *International Conference on Electronics, Information and Communications*, 2023. Cited on page 11.

[97] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, February 2014. arXiv:1312.6199. Cited on page 23.

[98] T. V. Nguyen, M. A. Dakka, S. M. Diakiw, M. D. VerMilyea, M. Perugini, J. M. M. Hall, and D. Perugini. A novel decentralized federated learning approach to train on globally

distributed, poor quality, and protected private medical data. *Scientific Reports*, 12(1), May 2022. Cited on page 11.

[99] Sam P. Tarassoli, Matthew E. Shield, Rhian S. Allen, Zita M. Jessop, Thomas D. Dobbs, and Iain S. Whitaker. Facial Reconstruction: A Systematic Review of Current Image Acquisition and Processing Techniques. *Frontiers in Surgery*, 7:537616, December 2020. Cited on page 5.

[100] U.S. Department of Health and Human Services. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.hhs.gov/hipaa/, 1996. Cited on page 4.

[101] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv:1706.03762. Cited on pages 16 and 17.

[102] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. Cited on page 32.

[103] Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. Cascaded Latent Diffusion Models for High-Resolution Chest X-ray Synthesis, March 2023. arXiv:2303.11224. Cited on page 17.

[104] Lilian Weng. From autoencoder to beta-vae. *lilianweng.github.io*, 2018. URL: https://lilianweng.github.io/posts/2018-08-12-vae/. Cited on page 7.

[105] Lilian Weng. What are diffusion models? *lilianweng.github.io*, Jul 2021. URL: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/. Cited on pages 6 and 8.

[106] Alexander Ziller, Dmitrii Usynin, Dmitrii Usynin, Rickmer Braren, Marcus R. Makowski, Marcus R. Makowski, Daniel Rueckert, and Georgios Kaissis. Medical imaging deep learning with differential privacy. *Scientific Reports*, 11(1):13524–13524, June 2021. Cited on page 12.